Fractal Geometry and Stochastics IV

BIRKHÄUSER

Christoph Bandt
Peter Mörters
Martina Zähle
Editors

Progress in Probability Vol. 61

BIRKHÄUSER

Progress in Probability
Volume 61

Fractal Geometry and Stochastics IV

Christoph Bandt
Peter Mörters
Martina Zähle
Editors

Birkhäuser
Basel · Boston · Berlin

Editors:

Christoph Bandt
Institut für Mathematik und Informatik
Ernst-Moritz-Arndt-Universität
17487 Greifswald
Germany
e-mail: bandt@uni-greifswald.de

Peter Mörters
Department of Mathematical Sciences
University of Bath
Bath BA2 7AY
UK
e-mail: maspm@bath.ac.uk

Martina Zähle
Mathematisches Institut
Friedrich-Schiller-Universität
07740 Jena
Germany
e-mail: martina.zaehle@uni-jena.de

2000 Mathematics Subject Classification: 20E08, 28A80, 35K05, 37B10, 37D20, 37E35, 60B15, 60D05, 60G15, 60G17, 60G18, 60G50, 60G51, 60G52, 60H15, 60J45, 60J50, 60J80, 60J85, 60K35, 60K37, 82B27

Library of Congress Control Number: 2009931383

Bibliographic information published by Die Deutsche Bibliothek.
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.ddb.de

ISBN 978-3-0346-0029-3 Birkhäuser Verlag AG, Basel · Boston · Berlin

Basel · Boston · Berlin
P.O. Box 133, CH-4010 Basel, Switzerland
Part of Springer Science+Business Media
Printed on acid-free paper produced from chlorine-free pulp. TCF ∞
Printed in Germany

ISBN 978-3-0346-0029-3 e-ISBN 978-3-0346-0030-9

9 8 7 6 5 4 3 2 1 www.birkhauser.ch

Contents

Part 1: Analysis on Fractals

Part 2: Conformal Dynamics and Schramm-Loewner Evolution

Part 3: Random Fractal Processes

Part 4: Emergence of Fractals in Complex Systems

Part 5: Iterated Function Schemes and Transformations of Fractals

Progress in Probability, Vol. 61, vii
© 2009 Birkhäuser Verlag Basel/Switzerland

Preface

The conference 'Fractal Geometry and Stochastics IV' with 116 participants from all over the world took place at Greifswald, Germany, from September 8–12, 2008. (More information can be found on the web page http://www.math-inf.uni-greifswald.de/∼bandt/fgs4.html.) The conference continued a series of meetings reflecting the developments in this modern mathematical field over the last fifteen years. As in the previous meetings – the former authors and their topics are listed at the end of the book – only the main speakers of the conference were asked for contributions to this volume. Again, we followed the principle to invite representatives of very active areas of research. Their articles give nice introductions to the subjects, most of them in form of surveys with selected proofs. Some of the contributions also contain interesting original results. The volume is addressed to non-experts as well as to specialists.

A few days before the conference Oded Schramm, outstanding researcher in areas closely related to fractal geometry and stochastics, died in a tragic hiking accident in the mountains near Seattle. The influence of his mathematical ideas was very strongly felt during the conference and is also present in the contributions in this volume. Some of our speakers were collaborators and friends of Oded and many of us benefited from his original ideas and clear insights in both conformal geometry and probability. His life was commemorated in a minute of silence during the opening talk by Greg Lawler, and some of our authors have dedicated their contributions to him.

We would like to express our gratitude to the Deutsche Forschungsgemeinschaft for the essential financial support for the conference and to the referees for their help in preparing this volume.

Progress in Probability, Vol. 61, ix–x
© 2009 Birkhäuser Verlag Basel/Switzerland

Introduction

The purpose of this book is to give insight into some exciting recent developments in the area of fractals, which are deeply influenced by the close interplay between geometry, analysis and probability, as well as by algebraic approaches as a new direction of research. In many of the topics the influence of randomness plays a major role. This reflects, in particular, natural relationships to models in statistical physics, mathematical genetics, population biology, finance and economics. For the convenience of the reader we have divided the book into five parts, corresponding to different directions of current research.

In Part 1 certain classes of fractals are treated in the more general frameworks of analysis on metric measure spaces and of self-similar algebraic groups acting on homogeneous rooted trees. The article of A. Grigor'yan, J. Hu and K.-S. Lau gives a survey on recent developments in the study of heat kernels on metric measure spaces. It explains several results obtained by the authors and their collaborators in a line of papers. In particular, it illustrates how heat kernel estimates of rather general form imply certain doubling properties of the underlying measure. Also relations to embeddings of the associated Dirichlet space into Besov type spaces and vice versa are shown. In addition a parabolic maximum principle is proved, along with some applications. Fractal sets with certain self-similarity properties serve as special examples. The latter may also be obtained as limit sets in the recently developed theory of self-similar groups. This leads, in particular, to an algebraic approach to Laplacians on such fractals. In the contribution of V. Kaimanovich a survey on new methods and results for self-similar groups is presented focusing on relationships between random walks on these structures and their amenability. Self-similar groups generated by bounded automata are discussed as a special case. The paper is completed by many references to the current literature on general self-similar groups.

Part 2 deals with a modern field in conformal dynamics. The Schramm-Loewner evolution (SLE) is a conformally invariant stochastic process consisting of a family of random planar curves. They are generated by solving Charles Loewner's differential equation with Brownian motion as input. SLE was discovered by Oded Schramm (2000) and developed by him together with Gregory Lawler and Wendelin Werner in a series of joint papers, for which they were awarded several prizes. SLE is conjectured or proved to be the scaling limit of various critical percolation models, and other stochastic processes in the plane with important applications in statistical physics. In the present book G.F. Lawler derives some new results: Using the Girsanov transformation and stochastic calculus he obtains large deviation (multifractal) estimates for a reverse flow associated with SLE. Moreover, novel fractal tools are developed to give a more accessible proof of Beffara's theorem (2008) on the Hausdorff dimension of the SLE curves.

In Part 3 some old and new relationships between fractal geometry and stochastic processes are reviewed by D. Khoshnevisan, where Lévy processes are of special interest. Connections to stochastic partial differential equations associated with the generator of such a process are also discussed. For proving geometric properties of occupation measure, range and local time Fourier transformation arguments are used as a main tool.

In Part 4 we show how recent developments in various areas of probability have led to the discovery and study of new fractal objects. The article of J. Steif surveys the model of dynamical percolation. Generically, this can be interpreted as a family of strongly coupled percolation processes indexed by a continuous time parameter. This setup allows to ask whether there exist exceptional times when the process has a property that has probability zero at any fixed time. In the cases described in this survey, the exceptional times form an interesting random fractal. The contribution of J. Blath describes a class of processes arising in mathematical genetics and population biology from the point of view of their fractal properties and, in particular, a new fractal phenomenon, the flickering of random measures, is described. G. Miermont deals with the classical probabilistic question of convergence of rescaled probabilistic objects to universal objects in the new context of random planar maps, such as random quadrangulations of a 2-sphere. The resulting limiting object, which is homeomorphic to the 2-sphere but has Hausdorff dimension 4, is again an exciting new fractal object.

Part 5 concerns (random) fractals generated by iterated function systems in an extended sense. M.F. Barnsley studies homeomorphisms between the corresponding attractors for equivalent address structures of the coding maps. In particular, the notion of fractal tops is discussed. Furthermore, generalized Minkowski metrics are considered which make affine iterated function systems hyperbolic. The paper of M. Furukado, S. Ito and H. Rao contains some new constructions and original results on the theory of Rauzy fractals based on the interplay between symbolic dynamics and domain-exchange transformations. The special class of (non-linear) Cantor sets is of traditional interest in fractal geometry and physical applications. F.M. Dekking answers in his article the question, whether the algebraic difference of two independent copies contains an interval or not, for two families of random Cantor sets. The survey is based on earlier papers by him, K. Simon and other coauthors.

Part 1

Analysis on Fractals

Progress in Probability, Vol. 61, 3–44

Heat Kernels on Metric Spaces with Doubling Measure

Alexander Grigor'yan, Jiaxin Hu and Ka-Sing Lau

Abstract. In this survey we discuss heat kernel estimates of self-similar type on metric spaces with doubling measures. We characterize the tail functions from heat kernel estimates in both non-local and local cases. In the local case we also specify the domain of the energy form as a certain Besov space, and identify the walk dimension in terms of the critical Besov exponent. The techniques used include self-improvement of heat kernel upper bound and the maximum principle for weak solutions. All proofs are completely analytic.

Mathematics Subject Classification (2000). Primary: 47D07,
Secondary: 28A80, 46E35.

Keywords. Doubling measure, heat kernel, maximum principle, heat equation.

1. Introduction

The heat kernel is an important tool in modern analysis, which appears to be useful for applications in mathematical physics, geometry, probability, fractal analysis, graphs, function spaces and in other fields. There has been a vast literature devoted to various aspects of heat kernels (see, for example, a collection [29]). It is not feasible to give a full-scale panorama of this subject here. In this article, we consider heat kernels on abstract metric measure spaces and focus on the following questions:

- Assuming that heat kernel satisfies certain estimates of self-similar type, what are the consequences for the underlying metric measure structure?
- Developing of the self-improvement techniques for heat kernel upper bounds of subgaussian types.

A.G. was supported by SFB 701 of the German Research Council (DFG) and the Grants from the Department of Mathematics and IMS in CUHK. J.H. was supported by NSFC (Grant No. 10631040) and the HKRGC Grant in CUHK. K.S.L was supported by the HKRGC Grant in CUHK.

Useful auxiliary tools that we develop here include the family of Besov function spaces and the maximum principle for weak solution for abstract heat equation.

Some of these questions have been discussed in various settings, for example, in [1, 4, 10, 12, 14, 18, 32, 33, 35, 36, 37, 38, 39] for the Euclidean spaces or Riemannian manifolds, in [5, 7, 25] for torus or infinite graphs, in [9, 27, 41] for metric spaces, in [2, 3, 6, 26] for certain classes of fractals. The contents of this paper are based on the work [20], [21], [22] and [24]. Similar questions were discussed in the survey [19] when the underlying measure is Ahlfors-regular, while the main emphasis in the present survey is on the case of doubling measures.

Notation. The sign $\asymp$ below means that the ratio of the two sides is bounded from above and below by positive constants. Besides, c is a positive constant, whose value may vary in the upper and lower bounds. The letters C, C', c, c' will always refer to positive constants, whose value is unimportant and may change at each occurrence.

2. What is a heat kernel

We give the definition of a heat kernel on a metric measure space, followed by some well-known examples on Riemannian manifolds and on a certain class of fractals.

2.1. The notion of a heat kernel

Let (M, d) be a locally compact separable metric space and let μ be a Radon measure on M with full support. The triple (M, d, μ) is termed a *metric measure space*. In the sequel, the norm in the real Banach space $L^p := L^p(M, \mu)$ is defined as usual by

$$\|f\|_p := \left(\int_M |f(x)|^p \, d\mu(x) \right)^{1/p}, \quad 1 \le p < \infty,$$

and

$$\|f\|_\infty := \operatorname*{esup}_{x \in M} |f(x)|,$$

where esup is the essential supremum. The inner product of $f, g \in L^2$ is denoted by (f, g).

Definition 2.1. A family $\{p_t\}_{t>0}$ of functions $p_t(x, y)$ on $M \times M$ is called a *heat kernel* if for any $t > 0$ it satisfies the following five conditions:

1. *Measurability:* the $p_t(\cdot, \cdot)$ is $\mu \times \mu$ measurable in $M \times M$.
2. *Markovian property:* $p_t(x, y) \ge 0$ for μ-almost all $x, y \in M$, and
$$\int_M p_t(x, y) d\mu(y) \le 1, \tag{2.1}$$
for μ-almost all $x \in M$.
3. *Symmetry:* $p_t(x, y) = p_t(y, x)$ for μ-almost all $x, y \in M$.

4. *Semigroup property:* for any $s > 0$ and for μ-almost all $x, y \in M$,

$$p_{t+s}(x,y) = \int_M p_t(x,z)p_s(z,y)d\mu(z). \tag{2.2}$$

5. *Approximation of identity*: for any $f \in L^2$,

$$\int_M p_t(x,y)\, f(y)\, d\mu(y) \overset{L^2}{\to} f(x) \ \text{ as } t \to 0+ .$$

We say that a heat kernel p_t is *stochastically complete* if equality takes place in (2.1), that is, for any $t > 0$,

$$\int_M p_t(x,y)d\mu(y) = 1 \quad \text{for } \mu\text{-almost all } x \in M.$$

Typically a heat kernel is associated with a Markov process $\left(\{X_t\}_{t\geq 0}, \{\mathbb{P}_x\}_{x\in M}\right)$ on M, so that $p_t(x,y)$ is the transition density of X_t, that is,

$$\mathbb{P}_x(X_t \in A) = \int_A p_t(x,y)\, d\mu(y)$$

for any Borel set $A \subset M$ (see Fig. 1).

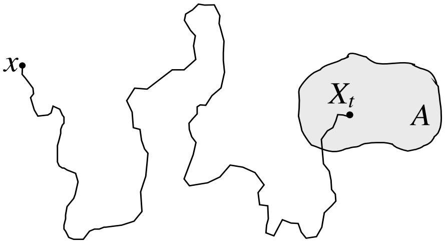

FIGURE 1. Markov process X_t hits the set A

Here are some examples of heat kernels.

Example 2.2. The best-known example of a heat kernel is the *Gauss-Weierstrass function* in $\mathbb{R}^n$:

$$p_t(x,y) = \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{|x-y|^2}{4t}\right). \tag{2.3}$$

It satisfies all the conditions of Definition 2.1 provided μ is the Lebesgue measure. This heat kernel is the transition density of the canonical Brownian motion in $\mathbb{R}^n$.

Example 2.3. The following function in $\mathbb{R}^n$

$$p_t(x,y) = \frac{C_n}{t^n}\left(1 + \frac{|x-y|^2}{t^2}\right)^{-\frac{n+1}{2}} \tag{2.4}$$

(where $C_n = \Gamma\left(\frac{n+1}{2}\right)/\pi^{(n+1)/2}$) is known on the one hand as *the Poisson kernel*, and on the other hand as the *density* of the *Cauchy distribution.* It is not difficult to verify that it also satisfies Definition 2.1 (also with respect to the Lebesgue measure) and, hence, is a heat kernel. The associated Markov process is the symmetric stable process of index 1.

More examples will be mentioned in the next section.

2.2. Heat semigroup and Dirichlet forms

The heat kernel is an integral kernel of a heat semigroup in L^2. A heat semigroup corresponds uniquely to a Dirichlet form in L^2.

A *Dirichlet form* $(\mathcal{E},\mathcal{F})$ in L^2 is a bilinear form $\mathcal{E}:\mathcal{F}\times\mathcal{F}\to\mathbb{R}$ defined on a dense subspace $\mathcal{F}$ of L^2, which satisfies in addition the following properties:

- *Positivity*: $\mathcal{E}(f):=\mathcal{E}(f,f)\geq 0$ for any $f\in\mathcal{F}$.
- *Closedness*: the space $\mathcal{F}$ is a Hilbert space with respect to the following inner product:
$$\mathcal{E}(f,g)+(f,g)\,.$$
- *The Markov property*: if $f\in\mathcal{F}$ then the function
$$g:=\min\{1,\max\{f,0\})\}$$
also belongs to $\mathcal{F}$ and $\mathcal{E}(g)\leq\mathcal{E}(f)$. Here we have used the shorthand $\mathcal{E}(f):=\mathcal{E}(f,f)$.

Any Dirichlet form has the *generator* $\mathcal{L}$, which is a non-positive definite self-adjoint operator on L^2 with domain $\mathcal{D}\subset\mathcal{F}$ such that
$$\mathcal{E}(f,g)=(-\mathcal{L}f,g)$$
for all $f\in\mathcal{D}$ and $g\in\mathcal{F}$. The generator determines the *heat semigroup* $\{P_t\}_{t\geq 0}$ defined by $P_t=e^{t\mathcal{L}}$. The heat semigroup satisfies the following properties:

- $\{P_t\}_{t\geq 0}$ is *contractive* in L^2, that is $\|P_tf\|_2\leq\|f\|_2$ for all $f\in L^2$ and $t>0$.
- $\{P_t\}_{t\geq 0}$ is *strongly continuous*, that is, for every $f\in L^2$,
$$P_tf\xrightarrow{L^2}f\text{ as }t\to 0+\,.$$
- $\{P_t\}_{t\geq 0}$ is *symmetric*, that is,
$$(P_tf,g)=(f,P_tg)\quad\text{for all }f,g\in L^2.$$
- $\{P_t\}_{t\geq 0}$ is *Markovian*, that is, for any $t>0$,
$$\text{if }f\geq 0\text{ then }P_tf\geq 0,\text{ and if }f\leq 1\text{ then }P_tf\leq 1.$$

Here and below the inequalities between L^2-functions are understood μ-almost everywhere in M.

The form $(\mathcal{E},\mathcal{F})$ can be recovered from the heat semigroup as follows. For any $t>0$, define a quadratic form $\mathcal{E}_t$ on L^2 as follows
$$\mathcal{E}_t(f):=\frac{1}{t}(f-P_tf,f)\,. \tag{2.5}$$
It is easy to show that $\mathcal{E}_t(f)$ is non-negative and is increasing as t is decreasing. In particular, it has the limit as $t\to 0$. It turns out that the limit is finite if and

only if $f \in \mathcal{F}$, and, moreover,

$$\lim_{t\to 0+} \mathcal{E}_t(f) = \mathcal{E}(f)$$

(cf. [9]). Extend $\mathcal{E}_t$ to a bilinear form as follows

$$\mathcal{E}_t(f,g) := \frac{1}{t}(f - P_t f, g).$$

Then, for all $f, g \in \mathcal{F}$,

$$\lim_{t\to 0+} \mathcal{E}_t(f,g) = \mathcal{E}(f,g).$$

The Markovian property of the heat semigroup implies that the operator P_t preserves the inequalities between functions, which allows to use monotone limits to extend P_t from L^2 to L^∞ and, in fact, to any L^q, $1 \le q \le \infty$. Moreover, the extended operator P_t is a contraction on any L^q (cf. [15, p.33]).

Recall some more terminology from the theory of the Dirichlet form (cf. [15]). The form $(\mathcal{E},\mathcal{F})$ is called *conservative* if $P_t 1 = 1$ for every $t > 0$. The form $(\mathcal{E},\mathcal{F})$ is called *local* if $\mathcal{E}(f,g) = 0$ for any couple $f, g \in \mathcal{F}$ with disjoint compact supports. The form $(\mathcal{E},\mathcal{F})$ is called *strongly local* if $\mathcal{E}(f,g) = 0$ for any couple $f, g \in \mathcal{F}$ with compact supports, such that $f \equiv \text{const}$ in an open neighborhood of $\operatorname{supp} g$.

The form $(\mathcal{E},\mathcal{F})$ is called *regular* if $\mathcal{F} \cap C_0(M)$ is dense both in $\mathcal{F}$ and in $C_0(M)$, where $C_0(M)$ is the space of all continuous functions with compact support in M, endowed with the sup-norm. For a non-empty open $\Omega \subset M$, let $\mathcal{F}(\Omega)$ be the closure of $\mathcal{F} \cap C_0(\Omega)$ in the norm of $\mathcal{F}$. It is known that if $(\mathcal{E},\mathcal{F})$ is regular, then $(\mathcal{E},\mathcal{F}(\Omega))$ is also a regular Dirichlet form in $L^2(\Omega,\mu)$.

Assume that the heat semigroup $\{P_t\}$ of a Dirichlet form $(\mathcal{E},\mathcal{F})$ in L^2 admits an integral kernel p_t, that is, for all $t > 0$ and $x \in M$, the function $p_t(x,\cdot)$ belongs to L^2, and the following identity holds:

$$P_t f(x) = \int_M p_t(x,y) f(y)\, d\mu(y), \tag{2.6}$$

for all $f \in L^2$ and μ-a.a. $x \in M$. Then the function p_t is indeed a heat kernel, as we will show below. For this reason, we also call p_t the heat kernel of the Dirichlet form $(\mathcal{E},\mathcal{F})$ or of the heat semigroup $\{P_t\}$.

Observe that if the heat kernel p_t of $(\mathcal{E},\mathcal{F})$ exists, then by (2.5) and (2.6),

$$\begin{align}\mathcal{E}_t(f) &= \frac{1}{2t}\int_M\int_M (f(y) - f(x))^2\, p_t(x,y) d\mu(y) d\mu(x) \tag{2.7}\\ &\quad + \frac{1}{t}\int_M (1 - P_t\mathbf{1}(x))\, f(x)^2\, d\mu(x) \tag{2.8}\end{align}$$

for any $t > 0$ and $f \in L^2$.

Proposition 2.4. ([21]) *If p_t is the integral kernel of the heat semigroup $\{P_t\}$, then p_t is a heat kernel.*

Proof. We will verify that p_t satisfies all the conditions in Definition 2.1. Let $t > 0$ be fixed until stated otherwise.

(1) Setting $p_{t,x} = p_t(x,\cdot)$, we see from (2.6) that, for any $f \in L^2$,

$$P_t f(x) = (p_{t,x}, f) \text{ for } \mu\text{-almost all } x \in M,$$

whence it follows that the function $x \mapsto (p_{t,x}, f)$ is μ-measurable in x. Let $\{\varphi_k\}_{k=1}^{\infty}$ be an orthonormal basis of L^2. Using the identity

$$p_t(x,y) = p_{t,x}(y) = \sum_{k=1}^{\infty} (p_{t,x}, \varphi_k)\, \varphi_k(y),$$

we conclude that $p_t(x,y)$ is jointly measurable in $x, y \in M$, because so are the functions $(p_{t,x}, \varphi_k)\, \varphi_k(y)$.

(2) By the Markovian property of P_t, for any non-negative function $f \in L^2$, there is a null set $\mathcal{N}_f \subset M$ such that

$$P_t f(x) \geq 0 \quad \text{for all } x \in M \setminus \mathcal{N}_f.$$

Let S be a countable family of non-negative functions, which is dense in the cone of all non-negative functions in L^2, and set

$$\mathcal{N} = \bigcup_{f \in S} \mathcal{N}_f$$

so that $\mathcal{N}$ is a null set. Then $P_t f(x) \geq 0$ for all $x \in M \setminus \mathcal{N}$ and for all $f \in S$. If f is any other non-negative function in L^2, then f is an L^2-limit of a sequence $\{f_k\} \subset S$, whence, for any $x \in M \setminus \mathcal{N}$,

$$(p_{t,x}, f) = \lim_{k\to\infty} (p_{t,x}, f_k) = \lim_{k\to\infty} P_t f_k(x) \geq 0.$$

Therefore, for any $x \in M \setminus \mathcal{N}$, we have that $p_{t,x} \geq 0$ μ-a.e. in M, which proves that $p_t(x,y) \geq 0$ for μ-a.a. $x, y \in M$.

Let $K \subset M$ be compact. Then the indicator function $\mathbf{1}_K$ belongs to L^2 and is bounded by 1, whence

$$\int_K p_t(x,y)\, d\mu(y) = P_t \mathbf{1}_K(x) \leq 1$$

for μ-a.a. $x \in M$. Choosing an increasing sequence of compact sets $\{K_n\}_{n=1}^{\infty}$ that exhausts M, we obtain that

$$\int_M p_t(x,y)\, d\mu(y) = \lim_{n\to\infty} \int_{K_n} p_t(x,y)\, d\mu \leq 1$$

for μ-a.a. $x \in M$.

Consequently, for any compact set $K \subset M$, we obtain by Fubini's theorem

$$\begin{aligned} \int_{K\times M} p_t(x,y)\, d\mu(y) d\mu(x) &= \int_K \left(\int_M p_t(x,y)\, d\mu(y) \right) d\mu(x) \\ &\leq \int_K d\mu(x) = \mu(K) < \infty, \end{aligned}$$

which implies that $p_t(x,y) \in L^1_{loc}(M \times M)$.

(3) For all $f, g \in L^2$, we have, again by Fubini's theorem,

$$(P_t f, g) = \int_{M \times M} p_t(x, y) f(y) g(x) d\mu(y) d\mu(x). \tag{2.9}$$

On the other hand, by the symmetry of P_t,

$$\begin{aligned}(P_t f, g) = (f, P_t g) &= \int_M P_t g(y) f(y) d\mu(y) \\ &= \int_{M \times M} p_t(y, x) f(y) g(x) d\mu(y) d\mu(x). \end{aligned} \tag{2.10}$$

Comparing (2.9) and (2.10), we obtain $p_t(x, y) = p_t(y, x)$ for μ-almost all $x, y \in M$.

(4) Using the semigroup identity $P_{t+s} = P_t(P_s)$ and Fubini's theorem, we obtain that, for any $f \in L^2$ and for μ-a.a. $x \in M$,

$$\begin{aligned} P_{t+s} f(x) &= P_t(P_s f)(x) \\ &= \int_M p_t(x, z) \left(\int_M p_s(z, y) f(y) d\mu(y) \right) d\mu(z) \\ &= \int_M \left(\int_M p_t(x, z) p_s(z, y) d\mu(z) \right) f(y) d\mu(y), \end{aligned}$$

whence, for any $g \in L^2$,

$$(P_{t+s} f, g) = \int_{M \times M} \left(\int_M p_t(x, z) p_s(z, y) d\mu(z) \right) f(y) g(x) d\mu(y) d\mu(x).$$

Comparing with

$$(P_{t+s} f, g) = \int_{M \times M} p_{t+s}(x, y) f(y) g(x) d\mu(y) d\mu(x),$$

we obtain (2.2).

(5) Finally, the approximation of identity property follows immediately from (2.6) and $P_t f \overset{L^2}{\to} f$ as $t \to 0$. □

Corollary 2.5. *If p_t and q_t are two integral kernels of a heat semigroup $\{P_t\}$, then, for any $t > 0$,*

$$p_t(x, y) = q_t(x, y) \text{ for } \mu\text{-a.a. } x, y \in M. \tag{2.11}$$

Proof. Similarly to (2.9), we have

$$(P_t f, g) = \int_{M \times M} q_t(x, y) f(y) g(x) d\mu(y) d\mu(x).$$

Comparing with (2.9), we obtain (2.11). □

Remark 2.6. Of course, not every heat semigroup possesses a heat kernel. The existence for the heat kernel and results related to these on-diagonal upper bounds can be found in [4, Theorem 2.1], [2, Propositions 4.13, 4.14], [8], [9], [11], [13], [14, Lemma 2.1.2], [17], [21], [23], [31], [40], [41], [42], [43].

2.3. Examples

Example 2.7. Let M be a connected Riemannian manifold, d be the geodesic distance, and μ be the Riemannian measure. The Laplace-Beltrami operator Δ on M can be made into a self-adjoint operator in $L^2(M,\mu)$ by appropriately defining its domain. Then Δ generates the heat semigroup $P_t = e^{t\Delta}$, which is associated with the local Dirichlet form $(\mathcal{E},\mathcal{F})$ where

$$\mathcal{E}(f) = \int_M |\nabla f|^2\, d\mu, \quad \mathcal{F} = W_0^{1,2}(M).$$

The corresponding Markov process is a Brownian motion on M.

It is known that this $\{P_t\}$ always has a smooth integral kernel $p_t(x,y)$, which is called the heat kernel of M. Although the explicit expression of $p_t(x,y)$ can not be given in general, there are many important classes of manifolds where $p_t(x,y)$ admits certain upper and/or lower bounds. For example, as it was proved in [32], if M is geodesically complete and its Ricci curvature is non-negative, then

$$p_t(x,y) \asymp \frac{1}{V\left(x,\sqrt{t}\right)} \exp\left(-c\frac{d(x,y)^2}{t}\right), \quad x,y \in M, t > 0, \tag{2.12}$$

where $V(x,r) = \mu\left(B(x,r)\right)$ is the volume of the geodesic ball

$$B(x,r) = \{y \in M : d(y,x) < r\}.$$

Example 2.8. If Δ is a self-adjoint Laplace operator as above then the operator $\mathcal{L} = -(-\Delta)^{\beta/2}$ (where $0 < \beta < 2$) generates on M a Markov process with jumps. In particular, if $M = \mathbb{R}^n$ then this is the symmetric stable process of index β, and the corresponding heat kernel admits the following estimate

$$p_t(x,y) \asymp \frac{1}{t^{n/\beta}} \left(1 + \frac{|x-y|^\beta}{t}\right)^{-\frac{n+\beta}{\beta}}.$$

A particular case $\beta = 1$ was already mentioned in Example 2.3.

Example 2.9. Let M be the Sierpinski gasket[1] in $\mathbb{R}^n$ (see Fig. 2).

It is known that the Hausdorff dimension of M is equal to $\alpha := \log(n+1)/\log 2$. Let μ be the α-dimensional Hausdorff measure on M, which clearly possesses the same self-similarity properties as the set M itself. It is possible to construct also a self-similar local Dirichlet form on M which possesses a continuous heat kernel, that is the transition density of a natural Brownian motion on M; moreover, the heat kernel admits the following estimate

$$p_t(x,y) \asymp \frac{1}{t^{\alpha/\beta}} \exp\left(-c\left(\frac{d(x,y)}{t^{1/\beta}}\right)^{\beta/(\beta-1)}\right), \tag{2.13}$$

where $\beta = \log(n+3)/\log 2$ is the *walk dimension* (see [6], [16], [30]). Similar results hold also for a large family of fractal sets, including p.c.f. fractals and the Sierpinski carpet in $\mathbb{R}^n$ (see [30] and [3]), but with different values of α and β.

[1] For the background of fractal sets including the notion of the Sierpinski gasket, see [2].

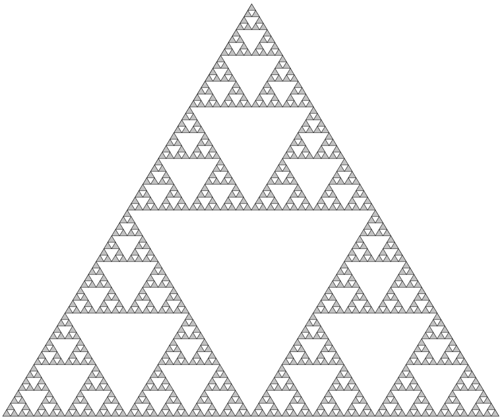

FIGURE 2. Sierpinski gasket in $\mathbb{R}^2$

3. Auxiliary material on metric measure spaces

Fix a metric measure space (M, d, μ) and define its volume function $V(x,r)$ by

$$V(x,r) := \mu(B(x,r))$$

where $x \in M$ and $r > 0$.

3.1. Besov spaces

Here we introduce function spaces $W^{\beta/2,p}$ on M. Choose parameters $1 \le p < \infty$, $\beta > 0$ and define the functional $E_{\beta,p}(u)$ for all functions $u \in L^p$ as follows:

$$E_{\beta,p}(u) = \sup_{0<r\le 1} r^{-p\beta/2} \int_M \left[\frac{1}{V(x,r)} \int_{B(x,r)} |u(y) - u(x)|^p \, d\mu(y) \right] d\mu(x). \tag{3.1}$$

For simplicity, if $p = 2$, denote it by

$$E_\beta(u) := E_{\beta,2}(u).$$

The *Besov* space $W^{\beta/2,p}$ is defined by

$$W^{\beta/2,p} := \{u \in L^p : \quad E_{\beta,p}(u) < \infty\} \tag{3.2}$$

with the norm

$$\|u\|_{W^{\beta/2,p}} := \|u\|_p + E_{\beta,p}(u)^{1/p}.$$

For Ahlfors regular[2] measures μ, the Besov space $W^{\beta/2,p}$ was introduced in [28, 34, 22] although using different notation.

It is not difficult to verify that for any $1 \le p < \infty$ and $\beta > 0$, the space $W^{\beta/2,p}$ is a Banach space. Note that the space $W^{\beta/2,p}$ decreases as β increases;

[2] A measure μ on a metric space (M, d) is said to be *Ahlfors-regular* if there exist $\alpha, c > 0$ such that $V(x,r) \asymp r^\alpha$ for all balls $B(x,r)$ in M with $r \in (0,1)$.

it may happen that this space becomes trivial for large enough β. For example, $W^{\beta/2,2}(\mathbb{R}^n) = \{0\}$ for $\beta > 2$.

Define the *critical Besov exponent* β^* by

$$\beta^* := \sup\left\{\beta > 0 : W^{\beta/2,2} \text{ is dense in } L^2(M,\mu)\right\}. \tag{3.3}$$

Lemma 3.1. *We have* $\beta^* \geq 2$.

Proof. It suffices to show that $W^{1,2}$ is dense in $L^2 = L^2(M,\mu)$. Let u be a Lipschitz function with a bounded support A and let A_r be the closed r-neighborhood of A. If L is the Lipschitz constant of u, then

$$\begin{aligned} E_2(u) &= \sup_{0<r\leq 1} r^{-2} \int_{A_r} \frac{1}{V(x,r)} \int_{B(x,r)} |u(y)-u(x)|^2 \, d\mu(y)d\mu(x) \\ &\leq \sup_{0<r\leq 1} r^{-2} \int_{A_r} L^2 r^2 d\mu(x) \\ &\leq L^2 \mu(A_1). \end{aligned}$$

It follows that $E_2(u) < \infty$ and hence $u \in W^{1,2}$. We are left to show that the class Lip of all Lipschitz functions with bounded supports is dense in L^2. Indeed, let now A be any bounded closed subset of M. For any positive integer n, consider the function on M

$$f_n(x) = (1 - nd(x,A))_+,$$

which is Lipschitz and is supported in $A_{1/n}$. Clearly, $f_n \to 1_A$ in L^2 as $n \to \infty$, whence it follows that $1_A \in \overline{\mathrm{Lip}}$, where the bar means the closure in L^2. Since the linear combinations of the indicator functions of bounded closed sets form a dense subset in L^2, it follows that $\overline{\mathrm{Lip}} = L^2$, which was to be proved. □

3.2. Doubling condition and reverse doubling condition

The measure μ on M is said to be *doubling* if there is a constant $C_D \geq 1$ such that

$$V(x,2r) \leq C_D V(x,r) \tag{3.4}$$

for all $x \in M$ and $r > 0$.

Proposition 3.2. *If* (3.4) *holds on* M, *then there exists* $\alpha > 0$ *depending only on the doubling constant* C_D *such that*

$$\frac{V(x,R)}{V(y,r)} \leq C_D \left(\frac{d(x,y)+R}{r}\right)^{\alpha} \quad \text{for all } x,y \in M \text{ and } 0 < r \leq R. \tag{VD}$$

Hence, the inequality of Proposition 3.2 can be used as an alternative definition of the doubling property of μ and will be referred to as (VD) (volume doubling). The advantage of this definition is that it introduces a parameter α that will frequently be used.

Proof. If $x = y$, then $R \leq 2^n r$ where

$$n = \left\lceil \log_2 \frac{R}{r} \right\rceil \leq \log_2 \frac{R}{r} + 1,$$

whence, it follows from (3.4) that

$$\frac{V(x,R)}{V(x,r)} \le \frac{V(x,2^n r)}{V(x,r)} \le (C_D)^n \le (C_D)^{\log_2 \frac{R}{r}+1} = C_D \left(\frac{R}{r}\right)^{\log_2 C_D}. \tag{3.5}$$

If $x \neq y$, then $B(x,R) \subset B(y,R+r_0)$ where $r_0 = d(x,y)$. By (3.5),

$$\frac{V(x,R)}{V(y,r)} \le \frac{V(y,R+r_0)}{V(y,r)} \le C_D \left(\frac{R+r_0}{r}\right)^{\log_2 C_D},$$

which finishes the proof. □

The measure μ satisfies a *reverse volume doubling* condition if there exist positive constants α' and c such that

$$\frac{V(x,R)}{V(x,r)} \ge c\left(\frac{R}{r}\right)^{\alpha'} \quad \text{for all } x \in M \text{ and } 0 < r \le R. \tag{RVD}$$

Proposition 3.3. *If* (M,d) *is connected and* μ *satisfies* (3.4), *then there exist positive constants* α' *and* c *such that* (RVD) *holds, provided* $B(x,R)^c$ *is non-empty.*

Proof. The condition $B(x,R)^c \neq \emptyset$ implies that

$$B(x,\rho') \setminus B(x,\rho) \neq \emptyset \tag{3.6}$$

for all $0 < \rho < R$ and $\rho' > \rho$. Indeed, otherwise M splits into disjoint union of two open sets: $B(x,\rho)$ and $\overline{B(x,\rho)}^c$. Since M is connected, the set $\overline{B(x,\rho)}^c$ must be empty, which contradicts the non-emptiness of $B(x,R)^c$.

If $0 < \rho \le R/2$, then we have by (3.6)

$$B\left(x,\frac{5}{3}\rho\right) \setminus B\left(x,\frac{4}{3}\rho\right) \neq \emptyset.$$

Let y be a point in this annulus. It follows from (VD) that

$$V(x,\rho) \le CV(y,\rho/3)$$

for some constant $C > 0$, whence

$$V(x,2\rho) \ge V(x,\rho) + V(y,\rho/3) \ge (1+\varepsilon)V(x,\rho), \tag{3.7}$$

where $\varepsilon = C^{-1}$.

For any $0 < r \le R$, we have that $2^n r \le R$ where

$$n := \left[\log_2 \frac{R}{r}\right] \ge \log_2 \frac{R}{r} - 1.$$

For any $0 \le k \le n-1$, we have $2^k r \le R/2$, and whence by (3.7),

$$V\left(x,2^{k+1}r\right) \ge (1+\varepsilon)V(x,2^k r).$$

Iterating this inequality, we obtain

$$\begin{aligned}\frac{V(x,R)}{V(x,r)} &\geq \frac{V(x,2^n r)}{V(x,r)} \geq (1+\varepsilon)^n \\ &\geq (1+\varepsilon)^{\log_2 \frac{R}{r}-1} = (1+\varepsilon)^{-1}\left(\frac{R}{r}\right)^{\log_2(1+\varepsilon)},\end{aligned}$$

thus proving (RVD). □

Remark 3.4. As one can see from the argument after (3.7), the measure μ is reverse doubling whenever the following inequality holds

$$V(x,Cr) \geq (1+\varepsilon)\,V(x,r) \tag{3.8}$$

for some $C>1$, $\varepsilon>0$ and all $x\in M$, $r>0$.

Corollary 3.5. *Assume that* (M,d) *is connected and* μ *satisfies* (VD)*. Then*

$$\mu(M)=\infty \ \Leftrightarrow\ \operatorname{diam}(M)=\infty \ \Leftrightarrow\ \text{(RVD)}.$$

Proof. If $\mu(M)=\infty$, then $\operatorname{diam}(M)=\infty$; indeed, otherwise M would be a ball of a finite radius and its measure would be finite by (VD). If $\operatorname{diam}(M)=\infty$, then $B^c(x,R)\neq\emptyset$ for any ball $B(x,R)$, and (RVD) holds by Proposition 3.3. Finally, (RVD) implies $\mu(M)=\infty$ by letting $R\to\infty$ in (RVD). □

4. Consequences of heat kernel estimates

We give here some consequences of the heat kernel estimates

$$\frac{1}{V\left(x,t^{1/\beta}\right)}\Phi_1\left(\frac{d(x,y)}{t^{1/\beta}}\right) \leq p_t(x,y) \leq \frac{1}{V\left(x,t^{1/\beta}\right)}\Phi_2\left(\frac{d(x,y)}{t^{1/\beta}}\right),$$

for all $t>0$ and μ-almost all $x,y\in M$. Functions $\Phi_1(s)$ and $\Phi_2(s)$ are always assumed to be non-negative and monotone decreasing on $[0,+\infty)$, the constant β is positive.

We prove that

- The lower estimate of the heat kernel implies that
  - the measure μ is doubling;
  - the space $\mathcal{F}$ is embedded in $W^{\beta/2,2}$;
  - the *lower tail function* $\Phi_1(s)$ is controlled from above by a negative power of s.
- The upper estimate of the heat kernel implies that
  - the space $W^{\beta/2,2}$ is embedded in $\mathcal{F}$;
  - if the Dirichlet form is non-local then the *upper tail function* $\Phi_2(s)$ is controlled from below by a negative power of s (for large s).

4.1. Consequences of lower bound

Let p_t be a heat kernel on a metric measure space (M, d, μ). Consider the lower estimate of p_t of the form:

$$p_t\left(x, y\right) \geq \frac{1}{V\left(x, t^{1/\beta}\right)} \Phi_1\left(\frac{d(x, y)}{t^{1/\beta}}\right) \tag{4.1}$$

for all $t > 0$ and μ-almost all $x, y \in M$.

Lemma 4.1. *Assume that the heat kernel p_t satisfies the lower bound* (4.1)*. If $\Phi_1(s_0) > 0$ for some $s_0 > 1$, then μ is doubling.*

Proof. Fix $r, t > 0$ and consider the following integral

$$\int_{B(x,r)} p_t(x, y) d\mu(y) := \int_M p_t(x, y) \mathbf{1}_{B(x,r)}(y)\, d\mu(y).$$

The right-hand side is understood as follows: the function

$$F(x, y) := p_t(x, y) \mathbf{1}_{B(x,r)}(y)$$

is measurable jointly in x, y so that, by Fubini's theorem, the integral

$$\int_M p_t(x, y) \mathbf{1}_{B(x,r)}(y)\, d\mu(y)$$

is well defined for μ-almost all $x \in M$ and is a measurable function of x. Choose any pointwise version of $p_t(x, y)$ as a function of x, y. By Fubini's theorem, there is a subset $M_0 \in M$ of full measure such that, for any $x \in M_0$, the function $p_t(x, y)$ is measurable in y and the inequalities (4.1) and (2.1) hold for μ-a.a. $y \in M$. It follows that, for all $x \in M_0$,

$$\int_{B(x,r)} p_t(x, y) d\mu(y) \leq 1 \tag{4.2}$$

whence

$$\frac{1}{V(x, r)} \geq \operatorname*{einf}_{y \in B(x,r)} p_t(x, y).$$

On the other hand, we have by (4.1)

$$\operatorname*{einf}_{y \in B(x,r)} p_t(x, y) \geq \frac{1}{V(x, t^{1/\beta})} \Phi_1\left(\frac{r}{t^{1/\beta}}\right),$$

which together with the previous estimate gives

$$\frac{V(x, r)}{V\left(x, t^{1/\beta}\right)} \leq \frac{1}{\Phi_1\left(r/t^{1/\beta}\right)}.$$

Setting here $t = \left(r/s_0\right)^\beta$ we obtain

$$\frac{V\left(x, r\right)}{V\left(x, r/s_0\right)} \leq \frac{1}{\Phi_1\left(s_0\right)}. \tag{4.3}$$

Since $s_0 > 1$ and $\Phi_1\left(s_0\right) > 0$, (4.3) implies that measure μ is doubling. □

Lemma 4.2. *Assume that the heat kernel p_t satisfies the lower bound* (4.1) *with $\Phi_1(s_0) > 0$ for some $s_0 \geq 1$. Then, there is a constant $c > 0$ such that for all $u \in L^2(M)$,*

$$\mathcal{E}(u) \geq c\, E_\beta(u). \tag{4.4}$$

Consequently, the space $\mathcal{F}$ embeds into $W^{\beta/2,2}$.

Proof. Let $t, r > 0$. It follows from (2.7) and the lower bound (4.1) that

$$\begin{aligned}\mathcal{E}(u) \geq \mathcal{E}_t(u) &\geq \frac{1}{2t}\int_M \int_{B(x,r)} (u(y)-u(x))^2\, p_t(x,y)d\mu(y)d\mu(x)\\ &\geq \frac{1}{2t}\Phi_1\left(\frac{r}{t^{1/\beta}}\right)\int_M \left(\frac{1}{V(x,t^{1/\beta})}\int_{B(x,r)} (u(y)-u(x))^2\, d\mu(y)\right) d\mu(x),\end{aligned}$$

where we have used the monotonicity of Φ_1. Choosing $t = (r/s_0)^\beta$ and noticing that $V(x, r/s_0) \leq V(x,r)$ by $s_0 \geq 1$, we obtain

$$\mathcal{E}(u) \geq \frac{s_0^\beta}{2r^\beta}\Phi_1(s_0)\int_M \left(\frac{1}{V(x,r)}\int_{B(x,r)} (u(y)-u(x))^2\, d\mu(y)\right) d\mu(x),$$

whence, by taking supremum in r,

$$\mathcal{E}(u) \geq \frac{1}{2}s_0^\beta \Phi_1(s_0) E_\beta(u),$$

thus proving (4.4). □

Finally, we give another consequences of the lower bound (4.1) of the heat kernel.

Lemma 4.3. *Assume that the heat kernel p_t satisfies the lower bound* (4.1)*. If μ satisfies the reverse doubling property* (RVD)*, then there is $c > 0$ such that*

$$\Phi_1(s) \leq c(1+s)^{-(\alpha'+\beta)} \quad \textit{for all } s > 0, \tag{4.5}$$

where α' is the same as in (RVD)*.*

Proof. Following [24], let $u \in L^2$ be a non-constant function. Choose a ball $B(x_0, R)$ where u is non-constant and let $a > b$ be two real values such that the sets

$$A = \{x \in B(x_0,R) : u(x) \geq a\} \quad \text{and} \quad B = \{x \in B(x_0,R) : u(x) \leq b\}$$

both have positive measure (see Fig. 3).
It follows from (2.7) that

$$\begin{aligned}\mathcal{E}(u) &\geq \frac{1}{2t}\int_M\int_M (u(y)-u(x))^2\, p_t(x,y)d\mu(y)d\mu(x)\\ &\geq \frac{1}{2t}\int_A\int_B (a-b)^2\,\frac{1}{V(x,t^{1/\beta})}\Phi_1\left(\frac{2R}{t^{1/\beta}}\right)d\mu(y)d\mu(x)\\ &= \frac{(a-b)^2}{2t}\Phi_1\left(\frac{2R}{t^{1/\beta}}\right)\mu(B)\int_A \frac{1}{V(x,t^{1/\beta})}d\mu(x).\end{aligned}$$

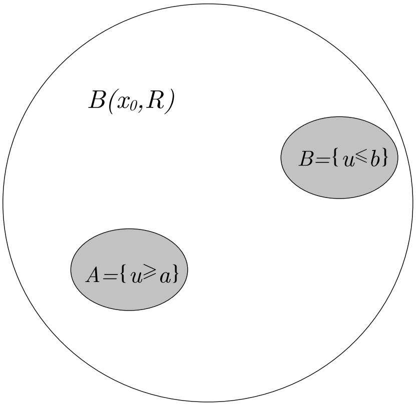

Figure 3. Sets A and B

For $x \in A$, we have that $B(x,R) \subset B(x_0,3R)$, and hence, for small enough $t>0$,

$$\frac{1}{V(x,t^{1/\beta})} = \frac{1}{V(x,R)} \cdot \frac{V(x,R)}{V(x,t^{1/\beta})} \geq \frac{1}{V(x_0,3R)} \cdot c\left(\frac{R}{t^{1/\beta}}\right)^{\alpha'},$$

where we have used the reverse doubling property (RVD). Therefore, for small $t>0$,

$$\mathcal{E}(u) \geq \frac{c'(a-b)^2}{V(x_0,3R)R^\beta}\mu(A)\mu(B)\left(\frac{2R}{t^{1/\beta}}\right)^{\alpha'+\beta}\Phi_1\left(\frac{2R}{t^{1/\beta}}\right).$$

If (4.5) fails, then there exists a sequence $\{s_k\}$ with $s_k \to \infty$ as $k \to \infty$ such that

$$s_k^{\alpha'+\beta}\Phi_1(s_k) \to \infty \quad \text{as} \quad k \to \infty.$$

Choose t_k such that $s_k = 2R/t_k^{1/\beta}$. Then

$$\left(\frac{2R}{t_k^{1/\beta}}\right)^{\alpha'+\beta}\Phi_1\left(\frac{2R}{t_k^{1/\beta}}\right) = s_k^{\alpha'+\beta}\Phi_1(s_k) \to \infty$$

as $k \to \infty$, and hence $\mathcal{E}(u) = \infty$. Hence, we see that $\mathcal{F}$ consists only of constants. Since $\mathcal{F}$ is dense in L^2, it follows that L^2 also consists of constants only. Hence, there is a point $x \in M$ with a positive mass, that is, $\mu(\{x\}) > 0$. Then (2.1) implies that, for all $t>0$,

$$p_t(x,x) \leq \frac{1}{\mu(\{x\})}. \tag{4.6}$$

However, by (RVD), we have $V(x,r) \to 0$ as $r \to 0$, which together with (4.1) implies that $p_t(x,x) \to \infty$ as $t \to 0$, thus contradicting (4.6). □

Remark 4.4. *The last argument in the above proof can be stated as follows. If* (RVD) *is satisfied and* (4.1) *holds with a function* Φ_1 *such that* $\Phi_1(0) > 0$*, then* $\mu(\{x\}) = 0$ *for all* $x \in M$*. This simple observation will also be used below.*

4.2. Consequences of upper bound

Consider the upper estimate of p_t of the form:

$$p_t(x,y) \le \frac{1}{V\left(x, t^{1/\beta}\right)} \Phi_2\left(\frac{d(x,y)}{t^{1/\beta}}\right) \tag{4.7}$$

for all $t > 0$ and μ-almost all $x, y \in M$.

Lemma 4.5. *Assume that* μ *satisfies both* (VD) *and* (RVD)*, and that the heat kernel* p_t *is stochastically complete and satisfies the upper bound* (4.7) *with*

$$\int_0^\infty s^{\alpha+\beta-1}\Phi_2(s)ds < \infty, \tag{4.8}$$

where α *is the same as in* (VD)*. Then, there is a constant* $c > 0$ *such that for all* $u \in L^2(M)$*,*

$$\mathcal{E}(u) \le CE_\beta(u). \tag{4.9}$$

Consequently, the space $W^{\beta/2,2}$ *embeds into* $\mathcal{F}$*.*

Proof. Fix $t \in (0,1)$ and let n be the smallest negative integer such that $2^{n+1} \ge t^{1/\beta}$. Since p_t is stochastically complete, we have that for any $t > 0$,

$$\mathcal{E}_t(u) = \frac{1}{2t}\int_M\int_M (u(x)-u(y))^2 p_t(x,y)d\mu(y)d\mu(x) = A_0(t)+A_1(t)+A_2(t) \tag{4.10}$$

where

$$A_0(t) := \frac{1}{2t}\int_M\int_{B(x,1)^c} (u(x)-u(y))^2 p_t(x,y)d\mu(y)d\mu(x), \tag{4.11}$$

$$A_1(t) := \frac{1}{2t}\int_M\int_{B(x,1)\setminus B(x,2^n)} (u(x)-u(y))^2 p_t(x,y)d\mu(y)d\mu(x), \tag{4.12}$$

$$A_2(t) := \frac{1}{2t}\int_M\int_{B(x,2^n)} (u(x)-u(y))^2 p_t(x,y)d\mu(y)d\mu(x). \tag{4.13}$$

Observing that by (VD)

$$\frac{V(x,2^{k+1})}{V(x,t^{1/\beta})} \le C\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^\alpha \text{ for all } k \ge n, \tag{4.14}$$

and using (4.7), we obtain

$$\begin{aligned}\int_{B(x,1)^c} p_t(x,y)d\mu(y) &\leq \sum_{k=0}^{\infty}\int_{B(x,2^{k+1})\setminus B(x,2^k)} \frac{1}{V(x,t^{1/\beta})}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)d\mu(y)\\ &\leq C\sum_{k=0}^{\infty}\frac{V(x,2^{k+1})}{V(x,t^{1/\beta})}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)\\ &\leq C'\sum_{k=0}^{\infty}\left(\frac{2^k}{t^{1/\beta}}\right)^{\alpha}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)\\ &\leq C'\int_{\frac{1}{2}t^{-1/\beta}}^{\infty} s^{\alpha-1}\Phi_2(s)ds \qquad (4.15)\\ &\leq ct\int_{\frac{1}{2}t^{-1/\beta}}^{\infty} s^{\alpha+\beta-1}\Phi_2(s)ds. \qquad (4.16)\end{aligned}$$

Applying the elementary inequality $(a-b)^2 \leq 2(a^2+b^2)$, we obtain from (4.11)

$$\begin{aligned}A_0(t) &\leq \frac{1}{t}\int_M\int_{B(x,1)^c}(u(x)^2+u(y)^2)p_t(x,y)d\mu(y)d\mu(x)\\ &= \frac{2}{t}\int_M u(x)^2\left(\int_{B(x,1)^c}p_t(x,y)d\mu(y)\right)d\mu(x)\\ &\leq 2c\|u\|_2^2\int_{\frac{1}{2}t^{-1/\beta}}^{\infty}s^{\alpha+\beta-1}\Phi_2(s)ds\\ &= o(1)\|u\|_2^2 \text{ as } t\to 0, \qquad (4.17)\end{aligned}$$

where we have used (4.8). It follows that

$$\lim_{t\to 0+} A_0(t) = 0. \qquad (4.18)$$

By (4.7) and (4.14), we obtain that, for $0 > k \geq n$,

$$\begin{aligned}&\int_{B(x,2^{k+1})\setminus B(x,2^k)}(u(x)-u(y))^2 p_t(x,y)d\mu(y)\\ &\leq \frac{1}{V(x,t^{1/\beta})}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)\int_{B(x,2^{k+1})\setminus B(x,2^k)}(u(x)-u(y))^2d\mu(y)\\ &\leq c\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^{\alpha}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)\frac{1}{V(x,2^{k+1})}\int_{B(x,2^{k+1})}(u(x)-u(y))^2d\mu(y).\end{aligned}$$

By the definition (3.1) of E_β, for all $k < 0$,

$$\int_M\frac{1}{V(x,2^{k+1})}\int_{B(x,2^{k+1})}(u(x)-u(y))^2d\mu(y)d\mu(x) \leq \left(2^{k+1}\right)^{\beta}E_\beta(u). \qquad (4.19)$$

Therefore, we obtain

$$
\begin{aligned}
A_1(t) &= \frac{1}{2t}\sum_{n\le k<0}\int_M\int_{B(x,2^{k+1})\setminus B(x,2^k)}(u(x)-u(y))^2p_t(x,y)d\mu(y)d\mu(x)\\
&\le \frac{1}{2t}\sum_{n\le k<0} c\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^{\alpha}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)\\
&\times\int_M\frac{1}{V(x,2^{k+1})}\int_{B(x,2^{k+1})}(u(x)-u(y))^2d\mu(y)d\mu(x) \qquad (4.20)\\
&\le c\sum_{n\le k<0}\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^{\alpha+\beta}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)E_\beta(u)\\
&\le cE_\beta(u)\int_0^\infty s^{\alpha+\beta-1}\Phi_2(s)ds, \qquad (4.21)
\end{aligned}
$$

where the latter integral converges due to (4.8).

For $k<n$, we have $2^{k+1}<t^{1/\beta}$ whence by (RVD)

$$
\frac{V(x,2^{k+1})}{V(x,t^{1/\beta})}\le c\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^{\alpha'}. \qquad (4.22)
$$

Similarly to the estimate of A_1, we obtain

$$
\begin{aligned}
A_2(t) &= \frac{1}{2t}\sum_{k<n}\int_M\int_{B(x,2^{k+1})\setminus B(x,2^k)}(u(x)-u(y))^2p_t(x,y)d\mu(y)d\mu(x)\\
&\le c\sum_{k<n}\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^{\alpha'+\beta}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)E_\beta(u)\\
&\le cE_\beta(u)\int_0^2 s^{\alpha'+\beta-1}\Phi_2(s)ds, \qquad (4.23)
\end{aligned}
$$

where the latter integral converges at 0 due to $\alpha'+\beta>0$. It follows from (4.10), (4.18), (4.21) and (4.23) that

$$
\mathcal{E}(u)=\lim_{t\to0+}\mathcal{E}_t(u)=\lim_{t\to0+}(A_0(t)+A_1(t)+A_2(t))\le CE_\beta(u),
$$

which finishes the proof. □

Lemma 4.6. *Assume that μ satisfies* (VD) *and that the heat kernel p_t satisfies the upper bound* (4.7)*. Then, either $(\mathcal{E},\mathcal{F})$ is local, or there is $c>0$ such that*

$$
\Phi_2(s)\ge c(1+s)^{-(\alpha+\beta)} \quad \textit{for all } s>0. \qquad (4.24)
$$

Proof. Let $u,v\in\mathcal{F}$ be functions with disjoint compact supports $A=\operatorname{supp}u$ and $B=\operatorname{supp}v$ (see Fig. 4).

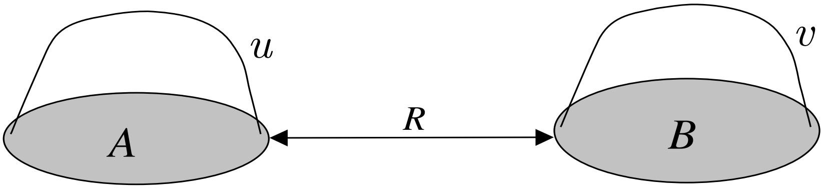

Figure 4. Functions u and v

Noticing that $(u,v)=0$, we obtain, for any $t>0$,

$$\begin{aligned}\mathcal{E}_t(u,v) &= \frac{1}{t}\,(u, v-P_t v)\\ &= -\frac{1}{t}\,(u, P_t v)\\ &= -\frac{1}{t}\int_A u(x)\left(\int_B v(y)p_t(x,y)d\mu(y)\right)d\mu(x).\end{aligned}$$

Setting $R=d(A,B)>0$ and using (4.7), we obtain

$$|\mathcal{E}_t(u,v)| \le \frac{1}{t}\Phi_2\left(\frac{R}{t^{1/\beta}}\right)\|v\|_1\int_A \frac{|u(x)|}{V(x,t^{1/\beta})}d\mu(x). \tag{4.25}$$

Choose any fixed point $x_0\in A$ and let $\operatorname{diam}(A)=r$. Then, using (VD), we see that, for all $x\in A$ and small $t>0$,

$$\begin{aligned}\frac{1}{V(x,t^{1/\beta})} &= \frac{1}{V(x_0,r)}\frac{V(x_0,r)}{V(x,t^{1/\beta})}\\ &\le \frac{c}{V(x_0,r)}\left(\frac{d(x_0,x)+r}{t^{1/\beta}}\right)^{\alpha} \le \frac{c}{V(x_0,r)}\left(\frac{2r}{t^{1/\beta}}\right)^{\alpha}.\end{aligned}$$

Therefore, by (4.25),

$$\begin{aligned}|\mathcal{E}_t(u,v)| &\le \frac{1}{t}\Phi_2\left(\frac{R}{t^{1/\beta}}\right)\|v\|_1\frac{c}{V(x_0,r)}\left(\frac{2r}{t^{1/\beta}}\right)^{\alpha}\|u\|_1\\ &= \frac{c(2r)^{\alpha}}{V(x_0,r)R^{\alpha+\beta}}\|u\|_1\|v\|_1\left(\frac{R}{t^{1/\beta}}\right)^{\alpha+\beta}\Phi_2\left(\frac{R}{t^{1/\beta}}\right).\end{aligned}$$

If (4.24) fails, then there exists a sequence $\{s_k\}$ such that $s_k\to\infty$ as $k\to\infty$, and

$$s_k^{\alpha+\beta}\Phi_2(s_k)\to 0.$$

Letting $t_k>0$ such that $s_k=R/t_k^{1/\beta}$, we obtain that

$$|\mathcal{E}_{t_k}(u,v)|\to 0 \quad \text{as} \quad k\to\infty,$$

showing that $\mathcal{E}(u,v)=0$. Hence, the $(\mathcal{E},\mathcal{F})$ is local, which was to be proved. □

4.3. Walk dimension

Here we obtain certain consequence of a two-sided estimate

$$\frac{1}{V\left(x,t^{1/\beta}\right)}\Phi_1\left(\frac{d(x,y)}{t^{1/\beta}}\right)\leq p_t\left(x,y\right)\leq\frac{1}{V\left(x,t^{1/\beta}\right)}\Phi_2\left(\frac{d(x,y)}{t^{1/\beta}}\right). \tag{4.26}$$

The parameter β from (4.26) is called the *walk dimension* of the associated Markov process.

Theorem 4.7. *Assume that μ satisfies both* (VD) *and* (RVD)*. Let the heat kernel $p_t\left(x,y\right)$ be stochastically complete and satisfy* (4.26) *where $\Phi_1\left(s_0\right)>0$ for some $s_0\geq 1$ and*

$$\int_0^{\infty}s^{\alpha+\beta+\varepsilon}\Phi_2(s)\frac{ds}{s}<\infty \tag{4.27}$$

for some $\varepsilon>0$. Then $\beta=\beta^*$ where β^* is the critical Besov exponent defined in (3.3)*.*

Remark 4.8. *Assuming in addition that $\Phi_1\left(s_0\right)>0$ for some $s_0>1$ allows to drop* (VD) *from the hypothesis, thanks to Lemma* 4.1*. If one assumes on top of that, that the metric space (M,d) is connected and has infinite diameter then* (RVD) *follows from* (VD) *by Corollary* 3.5*. Hence, in this case* (RVD) *can be dropped from the assumptions as well.*

Proof. By Lemma 4.2, we have the inclusion $\mathcal{F}\subset W^{\beta/2,2}$. Since $\mathcal{F}$ is always dense in L^2, we conclude that $W^{\beta/2,2}$ is dense in L^2, whence $\beta^*\geq\beta$.

To prove the opposite inequality, it suffices to verify that, for any $\beta'>\beta$, the space $W^{\beta'/2,2}$ is not dense in L^2. We can assume that $\beta'-\beta$ is sufficiently small so that the condition (4.27) holds with $\varepsilon=\beta'-\beta$.

Let us show that $u\in W^{\beta'/2,2}$ implies $\mathcal{E}\left(u\right)=0$. We use again the decomposition

$$\mathcal{E}_t\left(u\right)=A_0(t)+A_1(t)+A_2\left(t\right)$$

where $A_i(t)$ are defined in (4.11)–(4.13). As in the proof of Lemma 4.5, we have

$$\lim_{t\to 0}A_0\left(t\right)=0.$$

Let us estimate $A_1(t)$ similarly to the proof of Lemma 4.5 (and using the same notation), but use $E_{\beta'}$ instead of E_β. Indeed, using instead of (4.19) the inequality

$$\int_M\frac{1}{V(x,2^{k+1})}\int_{B(x,2^{k+1})}(u(x)-u(y))^2d\mu(y)d\mu\left(x\right)\leq\left(2^{k+1}\right)^{\beta'}E_{\beta'}\left(u\right), \tag{4.28}$$

we obtain from (4.20) that

$$\begin{aligned}A_1(t)&\leq ct^{\frac{\beta'}{\beta}-1}\sum_{n\leq k<0}\left(\frac{2^{k+1}}{t^{1/\beta}}\right)^{\alpha+\beta'}\Phi_2\left(\frac{2^k}{t^{1/\beta}}\right)E_{\beta'}(u)\\&\leq ct^{\frac{\beta'}{\beta}-1}E_{\beta'}(u)\int_0^{\infty}s^{\alpha+\beta'-1}\Phi_2(s)ds\end{aligned} \tag{4.29}$$

where the integral converges due to (4.27). In the same way, one obtains

$$A_2(t) \leq ct^{\frac{\beta'}{\beta}-1} E_{\beta'}(u) \int_0^2 s^{\alpha'+\beta'-1} \Phi_2(s) ds.$$

Putting together all the estimates, we obtain

$$\mathcal{E}_t(u) \leq A_0(t) + Ct^{\frac{\beta'}{\beta}-1} E_{\beta'}(u) \to 0 \quad \text{as } t \to 0,$$

whence

$$\mathcal{E}(u) = \lim_{t\to 0} \mathcal{E}_t(u) = 0.$$

Since $\mathcal{E}_t(u) \leq \mathcal{E}(u)$, this implies back that $\mathcal{E}_t(u) \equiv 0$ for all $t > 0$.

On the other hand, it follows from (2.7) and the lower bound in (4.26) that

$$\begin{aligned} \mathcal{E}_t(u) &\geq \frac{1}{2t} \iint\limits_{\left\{d(x,y)\leq s_0 t^{1/\beta}\right\}} (u(y)-u(x))^2 p_t(x,y) d\mu(y) d\mu(x) \\ &\geq \frac{\Phi_1(s_0)}{2t} \iint\limits_{\left\{d(x,y)\leq s_0 t^{1/\beta}\right\}} \frac{(u(x)-u(y))^2}{V\left(x,t^{1/\beta}\right)} d\mu(y) d\mu(x), \end{aligned}$$

which yields $u(x) = u(y)$ for μ-almost all x, y such that $d(x,y) \leq s_0 t^{1/\beta}$. Since t is arbitrary, we conclude that u is a constant function.

Hence, we have shown that the space $W^{\beta'/2,2}$ consists of constants. However, it follows from Remark 4.4 that the constant functions are not dense in L^2, which finishes the proof. □

4.4. Consequence of two-sided estimates (non-local case)

Lemmas 4.3 and 4.6 of the previous subsections imply immediately the following.

Theorem 4.9. *Assume that the metric measure space (M, d, μ) satisfies* (VD) *and* (RVD). *Let $\{p_t\}$ be a heat kernel on M such that, for all $t > 0$ and almost all $x, y \in M$,*

$$\frac{C_1'}{V\left(x,t^{1/\beta}\right)} \Phi\left(C_1 \frac{d(x,y)}{t^{1/\beta}}\right) \leq p_t(x,y) \leq \frac{C_2'}{V\left(x,t^{1/\beta}\right)} \Phi\left(C_2 \frac{d(x,y)}{t^{1/\beta}}\right) \tag{4.30}$$

where C_1, C_1', C_2, C_2' are positive constants. Then either the associated Dirichlet form $\mathcal{E}$ is local or

$$c_1(1+s)^{-(\alpha+\beta)} \leq \Phi(s) \leq c_2(1+s)^{-(\alpha'+\beta)} \tag{4.31}$$

for all $s > 0$ and some $c_1, c_2 > 0$, where α and α' are the exponents from (VD) *and* (RVD), *respectively.*

5. A maximum principle and its applications

5.1. Weak differentiation

Let $\mathcal{H}$ be a Hilbert space over $\mathbb{R}$ and I be an interval in $\mathbb{R}$. We say that a function $u : I \to \mathcal{H}$ is *weakly differentiable* at $t \in I$ if for any $\varphi \in \mathcal{H}$, the function $(u(\cdot), \varphi)$ is differentiable at t (where the outer brackets stand for the inner product in $\mathcal{H}$), that is, the limit

$$\lim_{\varepsilon \to 0} \left(\frac{u(t+\varepsilon) - u(t)}{\varepsilon}, \varphi \right)$$

exists. In this case it follows from the principle of uniform boundedness that there is $w \in \mathcal{H}$ such that

$$\lim_{\varepsilon \to 0} \left(\frac{u(t+\varepsilon) - u(t)}{\varepsilon}, \varphi \right) = (w, \varphi)$$

for all $\varphi \in \mathcal{H}$. We refer to the vector w as the *weak derivative* of the function u at t and write $w = u'(t)$. Of course, we have the weak convergence

$$\frac{u(t+\varepsilon) - u(t)}{\varepsilon} \rightharpoonup u'(t) \quad \text{as } \varepsilon \to 0.$$

In the next statement, we collect the necessary elementary properties of weak differentiation.

Lemma 5.1.

(i) *If u is weakly differentiable at t then u is strongly* (*that is, in the norm of $\mathcal{H}$*) *continuous at t.*

(ii) (The product rule) *If functions $u : I \to \mathcal{H}$ and $v : I \to \mathcal{H}$ are weakly differentiable at t, then the inner product (u, v) is also differentiable at t and*

$$(u, v)' = (u', v) + (u, v').$$

(iii) (The chain rule) *Let (M, μ) be a measure space and set $\mathcal{H} = L^2(M, \mu)$. Let $u : I \to L^2(M, \mu)$ be weakly differentiable at $t \in I$. Let Φ be a smooth real-valued function on $\mathbb{R}$ such that*

$$\Phi(0) = 0, \quad \sup_{\mathbb{R}} |\Phi'| < \infty, \; \sup_{\mathbb{R}} |\Phi''| < \infty. \tag{5.1}$$

Then the function $\Phi(u) : I \to L^2(M, \mu)$ is also weakly differentiable at t and

$$\Phi(u)' = \Phi'(u)\, u'.$$

Proof. To shorten the notation, we write u_t for $u(t)$.

(i) It suffices to verify that, for any sequence $\{\varepsilon_k\} \to 0$, we have

$$\|u_{t+\varepsilon_k} - u_t\| \to 0 \text{ as } k \to \infty. \tag{5.2}$$

The sequence $\frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}$ converges weakly, whence it follows that it is weakly bounded and, hence, also strongly bounded. The latter clearly implies (5.2).

(ii) Let $\{\varepsilon_k\}$ be as above. We have the identity

$$\frac{(u_{t+\varepsilon_k}, v_{t+\varepsilon_k}) - (u_t, v_t)}{\varepsilon_k} = \left(\frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}, v_t\right) + \left(u_t, \frac{v_{t+\varepsilon_k} - v_t}{\varepsilon_k}\right) + \left(\frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}, v_{t+\varepsilon_k} - v_t\right).$$

By the definition of the weak derivative, the first two terms in the right-hand side converge to (u'_t, v_t) and (u_t, v'_t) respectively. By part (i), the sequence $\frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}$ is bounded in norm, whereas $\|v_{t+\varepsilon_k} - v_t\| \to 0$ as $k \to \infty$; hence, the third term goes to 0, and we obtain the desired.

(iii) By (5.1) the function Φ admits the estimate $|\Phi(r)| \leq C|r|$ for all $r \in \mathbb{R}$, which implies that the function $\Phi(u_t)$ belongs to $L^2(M, \mu)$ for any $t \in I$. By the mean value theorem, for any $r, s \in \mathbb{R}$, there exists $\xi_{r,s} \in (0,1)$ such that

$$\Phi(r+s) - \Phi(r) = \Phi'(r + \xi_{r,s}(r-s))\, s.$$

We have then

$$\frac{\Phi(u_{t+\varepsilon_k}) - \Phi(u_t)}{\varepsilon_k} = \Phi'(u_t + \xi(u_{t+\varepsilon_k} - u_t)) \frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}$$

where we write for simplicity $\xi = \xi_{u_t, u_{t+\varepsilon_k} - u_t}$, which can also be rewritten as

$$\frac{\Phi(u_{t+\varepsilon_k}) - \Phi(u_t)}{\varepsilon_k} = \Big(\Phi'\big(u_t + \xi\left(u_{t+\varepsilon_k} - u_t\right)\big) - \Phi'(u_t)\Big) \frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k} + \Phi'(u_t) \frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}. \tag{5.3}$$

Since

$$\|\left(\Phi'(u_t + \xi(u_{t+\varepsilon_k} - u_t)) - \Phi'(u_t)\right)\| \leq \sup|\Phi''|\, \|u_{t+\varepsilon_k} - u_t\|$$

(where $\|\cdot\|$ is the L^2-norm) and the term $\frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}$ is bounded in norm, the first term in the right-hand side of (5.3) tends to 0 strongly as $k \to \infty$. We are left to verify that the second term goes weakly to $\Phi'(u_t)\, u'_t$. Indeed, for any $\varphi \in L^2(M, \mu)$, we have that

$$\left(\Phi'(u_t) \frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}, \varphi\right) = \left(\frac{u_{t+\varepsilon_k} - u_t}{\varepsilon_k}, \Phi'(u_t)\varphi\right) \to (u'_t, \Phi'(u_t)\varphi) = (\Phi'(u_t)u'_t, \varphi),$$

where we have used the fact that $\Phi'(u_t)$ is bounded and, hence, $\Phi'(u_t)\varphi \in L^2(M, \mu)$. □

5.2. Maximum principle for weak solutions

As before, let $(\mathcal{E}, \mathcal{F})$ be a regular Dirichlet form in $L^2(M, \mu)$. Consider a path $u : I \to \mathcal{F}$. We say that u is a *weak subsolution* of the heat equation in an open set $\Omega \subset M$ if u is weakly differentiable in the space $L^2(\Omega)$ at any $t \in I$ and, for any non-negative $\varphi \in \mathcal{F}(\Omega)$,

$$(u', \varphi) + \mathcal{E}(u, \varphi) \leq 0. \tag{5.4}$$

Similarly one defines the notions of weak supersolution and weak solution.

A similar definition was introduced in [20] but with the difference that the time derivative u' was understood in the sense of the norm convergence in $L^2(\Omega)$. Let us refer to the solutions defined in [20] as *semi-weak* solutions. Clearly, any semi-weak solution is also a weak solution.

It is easy to see that, for any $f \in L^2(M)$, the function $P_t f$ is a weak solution in $(0,\infty) \times \Omega$ for any open $\Omega \subset M$ (cf. [20, Example 4.10]).

Proposition 5.2 (parabolic maximum principle). *Let u be a weak subsolution of the heat equation in $(0,T) \times \Omega$, where $T \in (0,+\infty]$ and Ω is an open subset of M. Assume in addition that u satisfies the following boundary and initial conditions:*

- $u_+(t,\cdot) \in \mathcal{F}(\Omega)$ *for any* $t \in (0,T)$;
- $u_+(t,\cdot) \overset{L^2(\Omega)}{\longrightarrow} 0 \quad$ *as* $t \to 0$.

Then $u(t,x) \leq 0$ for any $t \in (0,T)$ and μ-almost all $x \in \Omega$.

Remark 5.3. For semi-weak solutions the maximum principle was proved in [20].

Remark 5.4. It was shown in [20, Lemma 4.4] that the condition $u_+ \in \mathcal{F}(\Omega)$ is equivalent to the following: $u \in \mathcal{F}$ and $u \leq v$ for some $v \in \mathcal{F}(\Omega)$. We will use this result to verify the boundary condition of the parabolic maximum principle.

Proof. Let Φ be a smooth function on $\mathbb{R}$ that satisfies the following conditions for some constant C:

(i) $\Phi(r) = 0$ for all $r \leq 0$;
(ii) $0 < \Phi'(r) \leq C$ for all $r > 0$.
(iii) $|\Phi''(r)| \leq C$ for all $r > 0$.

Then $\Phi(u) = \Phi(u_+) \in \mathcal{F}(\Omega)$ so that we can set $\varphi = \Phi(u)$ in (5.4) and obtain

$$(u', \Phi(u)) + \mathcal{E}(u, \Phi(u)) \leq 0.$$

Since Φ is increasing and Lipschitz, we conclude by [20, Lemma 4.3] that $\mathcal{E}(u, \Phi(u)) \geq 0$, whence it follows that

$$(u', \Phi(u)) \leq 0. \tag{5.5}$$

Since Φ satisfies the conditions (5.1), we conclude by Lemma 5.1, that the function $t \mapsto \Phi(u)$ is weakly differentiable in the space $L^2(\Omega)$ and

$$\Phi(u)' = \Phi'(u)\, u',$$

and $(u, \Phi(u))$ is differentiable in t and

$$\begin{aligned} (u, \Phi(u))' &= (u', \Phi(u)) + \left(u, \Phi(u)'\right) \\ &= (u', \Phi(u)) + (u, \Phi'(u)\, u') \\ &= (u', \Phi(u)) + (u', \Phi'(u)\, u). \end{aligned}$$

Set $\Psi(r) = \Phi'(r)\, r$ so that

$$(u, \Phi(u))' = (u', \Phi(u)) + (u', \Psi(u)).$$

Assume for a moment that function Ψ also satisfies the above properties (i)–(iii). Applying (5.5) to Ψ, we obtain from the previous line

$$(u, \Phi(u))' \leq 0,$$

that is, the function $t \mapsto (u, \Phi(u))$ is decreasing in t. By the properties (i)–(ii), we have $\Phi(r) \leq Cr$ for $r > 0$, which implies

$$(u, \Phi(u)) = (u_+, \Phi(u_+)) \leq C\|u_+\|^2 \to 0 \text{ as } t \to 0.$$

Hence, the function $t \mapsto (u_+, \Phi(u_+))$ is non-negative, decreasing and goes to 0 as $t \to 0$, which implies that this function is identical 0. It follows that $u_+ = 0$, which was to be proved.

We are left to specify the choice of Φ so that the function $\Psi(r) = \Phi'(r)\, r$ is also in the class (i)–(iii). Let us construct Φ from its derivative Φ' that can be chosen to satisfy the following:

- $\Phi'(r) = 0$ for $r \leq 0$;
- $\Phi'(r) = 1$ for $r \geq 1$;
- $\Phi''(r) > 0$ for $r \in (0, 1)$.

(see Fig. 5).

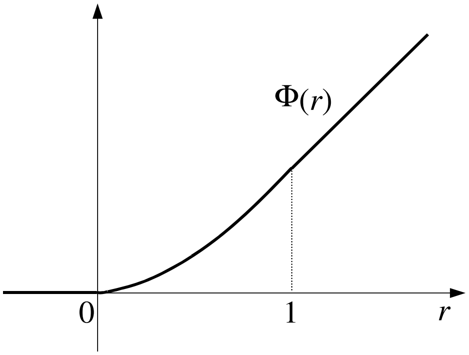

FIGURE 5. Function $\Phi(r)$

Clearly, Φ satisfies (i)–(iii). It follows from the identity

$$\Psi'(r) = \Phi''(r)\, r + \Phi'(r)$$

that $\Psi'(r)$ is bounded and $\Psi'(r) > 0$ for $r > 0$. Finally, we have

$$\Psi''(r) = \Phi'''(r)\, r + 2\Phi''(r)$$

whence it follows that $\Psi''(r) = 0$ for large enough r and, hence, Ψ'' is bounded. We conclude that Ψ satisfies (i)–(iii), which finishes the proof. □

5.3. Some applications of the maximum principle

Recall that if $(\mathcal{E}, \mathcal{F})$ is a regular Dirichlet form in $L^2(M, \mu)$ then, for any open set Ω, $(\mathcal{E}, \mathcal{F}(\Omega))$ is also a regular Dirichlet form in $L^2(\Omega, \mu)$. Denote by P_t^Ω the heat semigroup of $(\mathcal{E}, \mathcal{F}(\Omega))$.

Lemma 5.5. *Assume that $(\mathcal{E}, \mathcal{F})$ is a regular and local Dirichlet form. Let $u(t,x)$ be a weak subsolution of the heat equation in $(0,\infty) \times U$, where U is an open subset of M. Assume further, for any $t > 0$, $u(t,\cdot)$ is bounded in M and is non-negative in U. If*

$$u(t,\cdot) \overset{L^2(U)}{\longrightarrow} 0 \quad as \quad t \to 0 \tag{5.6}$$

then the following inequality holds for all $t > 0$ and almost all $x \in U$:

$$u(t,x) \le \left(1 - P_t^U \mathbf{1}_U(x)\right) \sup_{0<s\le t} \|u(s,\cdot)\|_{L^\infty(U)}. \tag{5.7}$$

Proof. We first assume that U is precompact. Choose an open set W such that $W \Subset U$. Fix a real $T > 0$ and set

$$m := \sup_{0<s\le T} \|u(s,\cdot)\|_{L^\infty(U)}. \tag{5.8}$$

We show that, for all $0 < t \le T$ and μ-almost all $x \in W$,

$$u(t,x) \le m\left(1 - P_t^W \mathbf{1}_W(x)\right). \tag{5.9}$$

Let ζ and η be cut-off functions[3] of the couples (W,U) and (U,M), respectively. Consider the function

$$w := \zeta u - m\left[\eta - P_t^W \mathbf{1}_W\right]. \tag{5.10}$$

Then (5.9) will follow if we prove that $w \le 0$ in $(0,T] \times W$.

Claim 1. The w is a weak subsolution of the heat equation in $(0,\infty) \times W$.

Clearly, $P_t^W \mathbf{1}_W$ is a weak solution of the heat equation in $(0,\infty) \times W$. Let us show that so is ζu. Indeed, the product ζu belongs to $\mathcal{F}$ because both ζ and u are in $L^\infty \cap \mathcal{F}$. For any test function $\psi \in \mathcal{F}(W)$, we have, using $\zeta\psi \equiv \psi$,

$$\begin{aligned}\left(\frac{\partial(\zeta u)}{\partial t}, \psi\right) &= \left(\zeta \frac{\partial}{\partial t} u, \psi\right) = \left(\frac{\partial}{\partial t} u, \psi\right) \\ &= -\mathcal{E}(u,\psi) = -\mathcal{E}(\zeta u, \psi) + \mathcal{E}((\zeta - 1)u, \psi) \\ &= -\mathcal{E}(\zeta u, \psi),\end{aligned}$$

where we have used also that $(\zeta - 1)u = 0$ in W and, hence,

$$\mathcal{E}((\zeta - 1)u, \psi) = 0,$$

by the locality of $(\mathcal{E}, \mathcal{F})$. Thus, ζu is a weak solution in $(0,\infty) \times W$.

[3] A cut-off function for the couple (W,U) is a function $\zeta \in \mathcal{F} \cap C_0(M)$ such that $0 \le \zeta \le 1$ in M, $\zeta = 1$ on an open neighborhood of $\overline{W}$, and $\operatorname{supp} \zeta \subset U$. If $(\mathcal{E}, \mathcal{F})$ is a regular Dirichlet form then a cut-off function exists for any couple (W,U) provided U is open and $\overline{W}$ is a compact subset of U (cf. [15, p. 27]).

Finally, the function $\eta(x)$ considered as a function of (t,x), is a weak supersolution of the heat equation in $(0,\infty)\times W$, since for any non-negative $\psi\in\mathcal{F}(W)$

$$\mathcal{E}(\eta,\psi)=\lim_{t\to 0}t^{-1}\left(\eta-P_t\eta,\psi\right)=\lim_{t\to 0}t^{-1}\left(1-P_t\eta,\psi\right)\geq 0,$$

whence it follows that w is a weak subsolution.

Claim 2. For every $t\in(0,T]$, we have $\left(w\left(t,\cdot\right)\right)_+\in\mathcal{F}(W)$.

By Remark 5.4, it suffices to prove that in $(0,T]\times M$

$$w\left(t,\cdot\right)\leq mP_t^W\mathbf{1}_W, \tag{5.11}$$

because $mP_t^W\mathbf{1}_W\in\mathcal{F}\left(W\right)$. In $M\setminus U$, inequality (5.11) holds trivially because

$$\zeta=0=P_t^W\mathbf{1}_W\quad\text{in }M\setminus U$$

and, hence, $w=-m\eta\leq 0$. To prove (5.11) in U, observe that $\eta=1$ in U and $0\leq u\leq m$ in $(0,T]\times U$, whence

$$w=\zeta u-m+mP_t^W\mathbf{1}_W\leq u-m+mP_t^W\mathbf{1}_W\leq mP_t^W\mathbf{1}_W,$$

which was to be proved.

Claim 3. The function w satisfies the initial condition

$$w\left(t,\cdot\right)\overset{L^2(W)}{\longrightarrow}0\text{ as }t\to 0. \tag{5.12}$$

Noticing that $\eta=1$ in W, we see that

$$\eta-P_t^W\mathbf{1}_W=\mathbf{1}_W-P_t^W\mathbf{1}_W\overset{L^2(W)}{\longrightarrow}0\text{ as }t\to 0.$$

Combining with (5.6), we obtain (5.12).

By the parabolic maximum principle (cf. Prop. 5.2), we obtain from Claims 1–3 that $w\leq 0$ in $(0,T]\times W$, thus proving (5.9).

Finally, let U be an arbitrary open subset of M. Let $\{W_i\}_{i=1}^\infty$ and $\{U_i\}_{i=1}^\infty$ be two increasing sequences of precompact open sets, both of which exhaust U, and such that $W_i\Subset U_i$ for all i. For each i, we have by (5.9) with $t=T$ that in W_i

$$u\leq\left[1-P_t^{W_i}\mathbf{1}_{W_i}\right]\sup_{0<s\leq t}\left\|u\left(s,\cdot\right)\right\|_{L^\infty(U_i)}. \tag{5.13}$$

Replacing by the monotonicity in the right-hand side U_i by U, and noticing that

$$P_t^{W_i}\mathbf{1}_{W_i}\overset{\text{a.e.}}{\longrightarrow}P_t^U\mathbf{1}_U\text{ as }i\to\infty,$$

we obtain (5.7) by letting $i\to\infty$ in (5.13). □

Corollary 5.6. *Assume that $(\mathcal{E},\mathcal{F})$ is regular and strongly local. Let $U\subset\Omega$ be two open subsets of M. Then the following inequality holds for all $t>0$ and μ-almost all $x\in U$:*

$$1-P_t^\Omega\mathbf{1}_\Omega(x)\leq\left(1-P_t^U\mathbf{1}_U(x)\right)\sup_{0<s\leq t}\left\|1-P_s^\Omega\mathbf{1}_\Omega\right\|_{L^\infty(U)}. \tag{5.14}$$

Proof. Approximating U by precompact open subsets, it suffices to prove the claim in the case when $U \Subset \Omega$. Let φ be a cut-off function of the couple (U, Ω). Then we can replace the term $1 - P_t^{\Omega}\mathbf{1}_{\Omega}(x)$ in the both sides of (5.14) by the function

$$u\,(t, x) = \varphi\,(x) - P_t^{\Omega}\mathbf{1}_{\Omega}(x).$$

Clearly, for any $t > 0$, the function $u\,(t, \cdot)$ is bounded in M, non-negative in U, and satisfies the initial condition (5.6). Let us verify that $u\,(t, x)$ is a weak solution of the heat equation in $(0, \infty) \times U$. It suffices to show that the function $\varphi\,(x)$ as a function of (t, x) is a weak solution in $(0, \infty) \times U$. Indeed, since φ is constant in a neighborhood of $\overline{U}$, the strong locality of $(\mathcal{E}, \mathcal{F})$ yields that $\mathcal{E}(\varphi, \psi) = 0$ for any $\psi \in \mathcal{F}(U)$, which finishes the proof. □

6. Upper bounds in the local case

6.1. Exponential tail

Following [21], we give an analytical approach of how to obtain the exponential tail of the heat kernel upper bound on the doubling space. This is a modification of the argument of [27]. For an alternative approach see [20] (and [25] for the case of infinite graphs).

The following is a key technical lemma.

Lemma 6.1. *Let $(\mathcal{E}, \mathcal{F})$ be a regular, strongly local Dirichlet form in $L^2(M, \mu)$. Let $\rho : [0, \infty) \to [0, \infty)$ be an increasing function. Assume that there exist $\varepsilon \in (0, \frac{1}{2})$ and $\delta > 0$ such that, for any ball $B \subset M$ of radius r and any positive t such that $\rho\,(t) \le \delta r$,*

$$P_t \mathbf{1}_{B^c} \le \varepsilon \ \text{ in } \frac{1}{4}B. \tag{6.1}$$

Then, for any $t > 0$ and any ball B of radius $r > 0$,

$$P_t \mathbf{1}_{B^c} \le C \exp\left(-c' t\, \Psi\left(\frac{cr}{t}\right)\right) \ \text{ in } \frac{1}{4}B, \tag{6.2}$$

where $C, c, c' > 0$ are constants depending on ε, δ, and function Ψ is defined by

$$\Psi(s) := \sup_{\lambda > 0} \left\{ \frac{s}{\rho(1/\lambda)} - \lambda \right\} \tag{6.3}$$

for all $s \ge 0$.

Remark 6.2. Letting $\lambda \to 0$ in (6.3), one sees that $\Psi\,(s) \ge 0$ for all $s \ge 0$. It is also obvious from (6.3) that $\Psi\,(s)$ is increasing in s.

Remark 6.3. If $\rho(t) = t^{1/\beta}$ for $\beta > 1$, then

$$\Psi(s) = \sup_{\lambda > 0} \left\{ s\lambda^{1/\beta} - \lambda \right\} = c_\beta s^{\beta/(\beta-1)}$$

for all $s \geq 0$, where $c_\beta > 0$ depends only on β (the supremum is attained for $\lambda = (s/\beta)^{\frac{\beta}{\beta-1}}$). The estimate (6.2) becomes

$$P_t \mathbf{1}_{B^c} \leq C \exp\left(-c\left(\frac{r^\beta}{t}\right)^{\frac{1}{\beta-1}}\right) \quad \text{in } \frac{1}{4}B.$$

Remark 6.4. If the heat semigroup P_t possesses the heat kernel $p_t(x, y)$ then the condition (6.1) can be equivalently stated as follows: If $\rho(t) \leq \delta r$ then, for almost all $x \in M$,

$$\int_{B(x,r)^c} p_t(x, y)\, d\mu(y) \leq \varepsilon. \tag{6.4}$$

Indeed, for any ball $B(x_0, r)$ and for almost all $x \in B(x_0, r/4)$, we have

$$P_t \mathbf{1}_{B(x_0,r)^c}(x) = \int_{B(x_0,r)^c} p_t(x, y)\, d\mu(y) \leq \int_{B(x,r/2)^c} p_t(x, y)\, d\mu(y),$$

so that (6.4) implies (6.1) (although with a different value of δ). Conversely, for almost all $x \in B(x_0, r/2)$,

$$\int_{B(x,r)^c} p_t(x, y)\, d\mu(y) \leq \int_{B(x_0,r/2)^c} p_t(x, y)\, d\mu(y) = P_t \mathbf{1}_{B(x_0,r/2)^c}(x),$$

so that (6.1) implies (6.4), for almost all $x \in B(x_0, r/8)$. Covering M by a countable family of balls of radius $r/8$, we obtain that (6.4) holds for almost all $x \in M$.

In the same way, the condition (6.2) is equivalent to the following: For all $\lambda, t, r > 0$ and for almost all $x \in M$,

$$\int_{B(x,r)^c} p_t(x, y)\, d\mu(y) \leq C \exp\left(-c' t\, \Psi\left(\frac{cr}{t}\right)\right). \tag{6.5}$$

Hence, Lemma 6.1 in the presence of the heat kernel can be stated as follows: If (6.4) holds for some $\varepsilon \in (0, 1/2)$, $\delta > 0$ and all $r, t > 0$ such that $\rho(t) \leq \delta r$ then (6.5) holds for all $r, t > 0$.

Proof of Lemma 6.1. Let us first show that the hypothesis (6.1) implies that there exist $\varepsilon \in (0, 1)$ and $\delta > 0$ such that, for any ball B of radius $r > 0$ and for any positive t such that $\rho(t) \leq \delta r$,

$$P_t^B \mathbf{1}_B \geq 1 - \varepsilon \quad \text{in } \frac{1}{4}B. \tag{6.6}$$

Indeed, applying [21, Proposition 4.7] we obtain that, for all t and almost everywhere in M,

$$P_t^B \mathbf{1}_{\frac{1}{2}B} \geq P_t \mathbf{1}_{\frac{1}{2}B} - \sup_{0<s\leq t} \left\| P_s \mathbf{1}_{\frac{1}{2}B} \right\|_{L^\infty\left(\left(\frac{3}{4}B\right)^c\right)}. \tag{6.7}$$

For any $x \in \frac{1}{4}B$, we have that $B(x, r/4) \subset \frac{1}{2}B$ (see Fig. 6).

Using the identity $P_t 1 = 1$ we obtain, for any $x \in \frac{1}{4}B$,

$$P_t \mathbf{1}_{\frac{1}{2}B} = 1 - P_t \mathbf{1}_{\left(\frac{1}{2}B\right)^c} \geq 1 - P_t \mathbf{1}_{B(x,r/4)^c}.$$

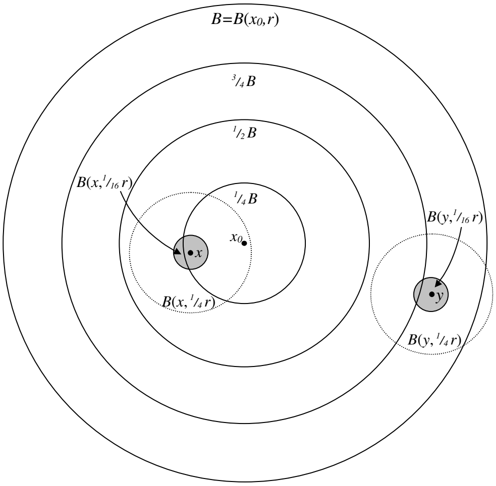

FIGURE 6. Illustration to the proof of (6.1)⇒(6.6)

Applying (6.1) for the ball $B(x, r/4)$, we obtain

$$P_t \mathbf{1}_{B(x,r/4)^c} \leq \varepsilon \text{ in } B(x, r/16),$$

provided that t satisfies

$$\rho(t) \leq \delta \frac{r}{4}. \tag{6.8}$$

It follows that, for any $x \in \frac{1}{4}B$,

$$P_t \mathbf{1}_{\frac{1}{2}B} \geq 1 - \varepsilon \text{ in } B(x, r/16),$$

whence

$$P_t \mathbf{1}_{\frac{1}{2}B} \geq 1 - \varepsilon \text{ in } \frac{1}{4}B. \tag{6.9}$$

On the other hand, for any $y \in \left(\frac{3}{4}\overline{B}\right)^c$, we have $\frac{1}{2}B \subset B(y, r/4)^c$ (see Fig. 6), whence

$$P_s \mathbf{1}_{\frac{1}{2}B} \leq P_s \mathbf{1}_{B(y,r/4)^c}.$$

Applying (6.1) for the ball $B(y, r/4)$ at time s, we obtain if (6.8) holds then, for all $0 < s \leq t$,

$$P_s \mathbf{1}_{B(y,r/4)^c} \leq \varepsilon \text{ in } B(y, r/16).$$

It follows that, for any $y \in \left(\frac{3}{4}\overline{B}\right)^c$,

$$P_s \mathbf{1}_{\frac{1}{2}B} \leq \varepsilon \text{ in } B(y, r/16),$$

whence

$$P_s \mathbf{1}_{\frac{1}{2}B} \leq \varepsilon \text{ in } \left(\frac{3}{4}\overline{B}\right)^c. \tag{6.10}$$

Combining (6.7), (6.9) and (6.10), we obtain that, under condition (6.8),

$$P_t^B \mathbf{1}_B \geq P_t^B \mathbf{1}_{\frac{1}{2}B} \geq 1 - 2\varepsilon \ \text{ in } \frac{1}{4}B, \tag{6.11}$$

which is equivalent to (6.6).

Now we show that (6.6) implies (6.2). The proof will be split into 5 steps.

Step 1. Assuming that

$$\rho\left(t\right) \leq \delta r \tag{6.12}$$

and that B is a ball of radius r, rewrite (6.6) in the form

$$1 - P_t^B \mathbf{1}_B \leq \varepsilon \ \text{ in } \frac{1}{4}B. \tag{6.13}$$

For any positive integer k, set $B_k = kB$ and we will prove that

$$1 - P_t^{B_k} \mathbf{1}_{B_k} \leq \varepsilon^k \ \text{ in } \frac{1}{4}B. \tag{6.14}$$

Since M is separable, there is a countable dense set of points in B_k. Let $\{b_j\}$ be a sequence of balls of radii r centered at those points. Clearly, $b_j \subset B_{k+1}$ and the family $\left\{\frac{1}{4}b_j\right\}$ covers B_k (see Fig. 7).

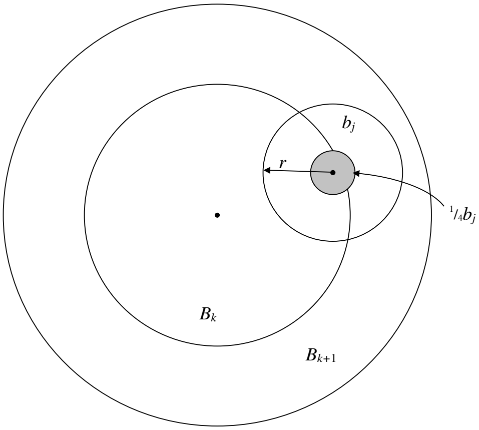

Figure 7. Balls $\{B_k\}$ and $\{b_j\}$

Due to (6.12), inequality (6.6) is valid for any ball b_j, that is, for all $0 < s \leq t$,

$$P_s^{B_{k+1}} \mathbf{1}_{B_{k+1}} \geq P_s^{b_j} \mathbf{1}_{b_j} \geq 1 - \varepsilon \text{ in } \frac{1}{4}b_j.$$

It follows that

$$P_s^{B_{k+1}} \mathbf{1}_{B_{k+1}} \geq 1 - \varepsilon \ \text{ in } B_k.$$

Applying the inequality (5.14) of Corollary 5.6 with $\Omega = B_{k+1}$ and $U = B_k$, we obtain that the following inequality holds in B_k:

$$\begin{aligned} 1 - P_t^{B_{k+1}} \mathbf{1}_{B_{k+1}} &\le \left(1 - P_t^{B_k} \mathbf{1}_{B_k}\right) \sup_{0<s\le t} \left\| 1 - P_s^{B_{k+1}} \mathbf{1}_{B_{k+1}} \right\|_{L^\infty(B_k)} \\ &\le \varepsilon \left(1 - P_t^{B_k} \mathbf{1}_{B_k}\right). \end{aligned}$$

Iterating in k and using (6.13), we obtain (6.14).

It follows from (6.14) that

$$P_t \mathbf{1}_{B_k^c} \le 1 - P_t \mathbf{1}_{B_k} \le 1 - P_t^{B_k} \mathbf{1}_{B_k} \le \varepsilon^k \text{ in } \frac{1}{4} B. \tag{6.15}$$

Although (6.15) has been proved for any integer $k \ge 1$, it is trivially true also for $k = 0$, if we define $B_0 := \emptyset$.

Step 2. Fix $t > 0$, $x \in M$ and consider the function

$$E_{t,x} = \exp\left(c \frac{d(x,\cdot)}{\rho(t)}\right), \tag{6.16}$$

where the constant $c > 0$ is to be determined later on. Set

$$r = \delta^{-1} \rho(t),$$

and we will prove that

$$P_t (E_{t,x}) \le C \text{ in } B(x, r/4), \tag{6.17}$$

where C is a constant depending on ε, δ. Set as before $B_k = B(x, kr)$, $k \ge 1$, and $B_0 = \emptyset$. Using (6.16) and (6.15), we obtain that in $B(x, r/4)$,

$$\begin{aligned} P_t (E_{t,x}) &= \sum_{k=0}^{\infty} P_t \left(\mathbf{1}_{B_{k+1} \setminus B_k} E_{t,x}\right) \\ &\le \sum_{k=0}^{\infty} \|E_{t,x}\|_{L^\infty(B_{k+1})} P_t \left(\mathbf{1}_{B_{k+1} \setminus B_k}\right) \\ &\le \sum_{k=0}^{\infty} \exp\left(c \frac{(k+1) r}{\rho(t)}\right) P_t \left(\mathbf{1}_{B_k^c}\right) \\ &\le \sum_{k=0}^{\infty} \exp\left(c(k+1)\delta^{-1}\right) \varepsilon^k. \end{aligned}$$

Choosing $c < \delta \log \frac{1}{\varepsilon}$ we obtain that this series converges, which proves (6.17).

Step 3. Let us prove that, for all $t > 0$ and $x \in M$,

$$P_t E_{t,x} \le C_1 E_{t,x}, \tag{6.18}$$

for some constant $C_1 = C(\varepsilon, \delta)$. Observe first that, for all $y, z \in M$, we have by the triangle inequality

$$\begin{aligned} E_{t,x}(y) &= \exp\left(c\frac{d(x,y)}{\rho(t)}\right) \\ &\le \exp\left(c\frac{d(x,z)}{\rho(t)}\right)\exp\left(c\frac{d(z,y)}{\rho(t)}\right) = E_{t,x}(z)E_{t,z}(y), \end{aligned}$$

which can also be written in the form of a function inequality:

$$E_{t,x} \le E_{t,x}(z)\, E_{t,z}.$$

It follows that

$$P_t(E_{t,x}) \le E_{t,x}(z)\, P_t(E_{t,z}). \tag{6.19}$$

By the previous step, we have

$$P_t(E_{t,z}) \le C \text{ in } B(z,r), \tag{6.20}$$

where $r = \frac{1}{4}\delta^{-1}\rho(t)$. For all $y \in B(z,r)$, we have

$$E_{t,z}(y) \le \exp\left(\frac{cr}{\rho(t)}\right) = \exp\left(c\delta^{-1}/4\right) =: C',$$

whence

$$E_{t,x}(z) \le E_{t,x}(y)\, E_{t,z}(y) \le C' E_{t,x}(y).$$

Letting y vary, we can write

$$E_{t,x}(z) \le C' E_{t,x} \text{ in } B(z,r).$$

Combining this with (6.19) and (6.20), we obtain

$$P_t(E_{t,x}) \le CC' E_{t,x} \text{ in } B(z,r).$$

Since z is arbitrary, covering M by a countable sequence of balls like $B(z,r)$, we obtain that (6.18) holds on M with $C_1 = CC'$.

Step 4. Let us prove that, for all $t > 0$, $x \in M$, and for any positive integer k,

$$P_{kt}(E_{t,x}) \le C_1^k \text{ in } \frac{1}{4}B, \tag{6.21}$$

where $B = \left(x, \delta^{-1}\rho(t)\right)$. Indeed, by (6.18)

$$P_{kt}(E_{t,x}) = P_{(k-1)t}P_t(E_{t,x}) \le C_1 P_{(k-1)t}E_{t,x}$$

which implies by iteration that

$$P_{kt}(E_{t,x}) \le C_1^{k-1} P_t E_{t,x}.$$

Combining with (6.17) and noticing that $C \le C_1$, we obtain (6.21).

Step 5. Fix a ball $B = B(x_0, r)$ and observe that (6.2) is equivalent to the following: for all $t, \lambda > 0$,

$$P_t \mathbf{1}_{B^c} \le C \exp\left(c'\lambda t - \frac{cr}{\rho(1/\lambda)}\right) \text{ in } \frac{1}{2}B, \tag{6.22}$$

where $C, c, c' > 0$ are constants depending on ε, δ. In what follows, we fix also t and λ.

Observe first that, for any $x \in \frac{1}{2}B$,

$$P_t \mathbf{1}_{B^c} \leq P_t \mathbf{1}_{B(x,r/2)^c}.$$

Hence, it suffices to prove that, for any $x \in \frac{1}{2}B$,

$$P_t \mathbf{1}_{B(x,r/2)^c} \leq C \exp\left(c'\lambda t - \frac{cr}{\rho(1/\lambda)}\right) \tag{6.23}$$

in a (small) ball around x. Covering then $\frac{1}{2}B$ by a countable family of such balls, we then obtain (6.22).

Changing t to t/k in (6.21), we obtain that

$$P_t\left(E_{t/k,x}\right) \leq C_1^k \text{ in } B\left(x, \sigma_k\right)$$

where $\sigma_k = \frac{1}{4}\delta^{-1}\rho\left(t/k\right)$. Since

$$E_{t/k,x} \geq \exp\left(c\frac{r}{\rho\left(t/k\right)}\right) \text{ in } B\left(x,r\right)^c$$

and, hence,

$$\mathbf{1}_{B(x,r)^c} \leq \exp\left(-\frac{cr}{\rho(t/k)}\right) E_{t/k,x},$$

we obtain that the following inequality holds in $B\left(x, \sigma_k\right)$

$$P_t \mathbf{1}_{B(x,r)^c} \leq \exp\left(-\frac{cr}{\rho(t/k)}\right) P_t\left(E_{t/k,x}\right) \leq \exp\left(c'k - \frac{cr}{\rho(t/k)}\right)$$

where $c' = \log C_1$. Given $\lambda > 0$, choose an integer $k \geq 1$ such that

$$\frac{k-1}{t} < \lambda \leq \frac{k}{t}.$$

Then we obtain the following inequality in $B\left(x, \sigma_k\right)$

$$P_t \mathbf{1}_{B(x,r)^c} \leq \exp\left(c'\left(\lambda t + 1\right) - \frac{cr}{\rho\left(1/\lambda\right)}\right), \tag{6.24}$$

which finishes the proof. □

6.2. Consequences of two-sided estimates (local case)

Now we are able to specify the local case in the statement of Theorem 4.9.

Given two points $x, y \in M$, a *chain* connecting x and y is any finite sequence $\{x_k\}_{k=0}^n$ of points in M such that $x_0 = x$, $x_n = y$. We say that a metric space satisfies the *chain condition* if there is a constant $C > 0$ such that for any positive integer n and for all $x, y \in M$ there is a chain $\{x_k\}_{k=0}^n$ connecting x and y, such that

$$d\left(x_k, x_{k+1}\right) \leq C\frac{d\left(x,y\right)}{n} \text{ for all } k = 0, 1, \ldots, n-1. \tag{6.25}$$

For example, the geodesic distance on any length space satisfies the chain condition. On the other hand, the combinatorial distance on a graph does not satisfy it.

Theorem 6.5. *Assume that the metric measure space* (M, d, μ) *satisfies the chain condition and that* μ *satisfies* (VD) *and* (RVD)*. Let* $(\mathcal{E}, \mathcal{F})$ *be a regular, local and conservative Dirichlet form, and let* $\{p_t\}$ *be the associated heat kernel such that, for all* $t > 0$ *and almost all* $x, y \in M$,

$$\frac{C_1'}{V\left(x, t^{1/\beta}\right)} \Phi\left(C_1 \frac{d\left(x, y\right)}{t^{1/\beta}}\right) \leq p_t\left(x, y\right) \leq \frac{C_2'}{V\left(x, t^{1/\beta}\right)} \Phi\left(C_2 \frac{d\left(x, y\right)}{t^{1/\beta}}\right) \tag{6.26}$$

where C_1, C_1', C_2, C_2' *are positive constants,* α, α' *are the exponents from* (VD) *and* (RVD)*, respectively, and* $\beta > \alpha - \alpha'$*. Then* $\beta \geq 2$ *and the following inequality holds:*

$$c_1' \exp\left(-c_1 s^{\beta/(\beta-1)}\right) \leq \Phi(s) \leq c_2' \exp\left(-c_2 s^{\beta/(\beta-1)}\right) \tag{6.27}$$

for some positive constants c_1, c_1', c_2, c_2' *and all* $s > 0$.

Proof. Let us first observe that the locality and the conservativeness imply the strong locality. Indeed, by [15, Lemma 4.5.2, p. 159 and Lemma, p. 161], we have the following identity

$$\lim_{t \to 0} \frac{1}{t} \int_M \left(1 - P_t 1\right) u^2 \, d\mu = \int_M \widetilde{u}^2 dk$$

for any $u \in \mathcal{F}$ where k is the killing measure of $(\mathcal{E}, \mathcal{F})$ and $\widetilde{u}$ is a quasi-continuous version of u. Since $P_t 1 = 1$, it follows that $k = 0$. Therefore, by the Beurling-Deny formula [15, Theorem 3.2.1, p. 108], $(\mathcal{E}, \mathcal{F})$ is strongly local. This will allow us to apply later Lemma 6.1.

We split the further proof into five steps.

Step 1. By Lemma 4.3 and the lower bound of p_t, we obtain

$$\Phi(s) \leq c(1+s)^{-(\alpha'+\beta)} \quad \text{for all } s > 0.$$

Therefore, using the upper bound of p_t, we obtain (similar to (4.15)) that

$$\begin{aligned} \int_{B(x,r)^c} p_t(x,y) d\mu(y) &\leq c \int_{\frac{1}{2}r/t^{1/\beta}}^{\infty} s^{\alpha-1} \Phi(s) ds \\ &\leq c' \int_{\frac{1}{2}r/t^{1/\beta}}^{\infty} s^{\alpha-\alpha'-\beta-1} ds. \end{aligned}$$

Due to the condition $\beta > \alpha - \alpha'$, the integral in the right-hand side converges and, hence, the right-hand side can be made arbitrarily small provided $rt^{-1/\beta}$ is large enough. We conclude by Lemma 6.1 (cf. Remark 6.4) with $\rho(t) = t^{1/\beta}$ that, for all $r, t > 0$ and for almost all $x \in M$,

$$\int_{B(x,r)^c} p_t\left(x, y\right) d\mu\left(y\right) \leq C \exp\left(-c' t \, \Psi\left(\frac{cr}{t}\right)\right), \tag{6.28}$$

where

$$\Psi(s) := \sup_{\lambda > 0} \left\{ s \lambda^{1/\beta} - \lambda \right\}. \tag{6.29}$$

Step 2. Let us prove that $\beta > 1$. If $\beta < 1$ then it follows from (6.29) that $\Psi \equiv \infty$. Substituting into (6.28) and letting $r \to 0$, we obtain that, for almost all $x \in M$,

$$\int_{M \setminus \{x\}} p_t(x,y)\, d\mu(y) = 0.$$

It follows from the stochastic completeness that there is a point $x \in M$ of a positive measure, which contradicts Remark 4.4.

Assume now that $\beta = 1$. Then by (6.29)

$$\Phi(s) = \begin{cases} 0, & 0 \le s \le 1 \\ \infty, & s > 1, \end{cases}$$

which implies that, for all $t < cr$ and for almost all $x \in M$,

$$\int_{B(x,r)^c} p_t(x,y)\, d\mu(y) = 0,$$

that is, $p_t(x,y) = 0$ for all $t < cd(x,y)$ and almost all $x, y \in M$. Together with (6.26), this yields the following bounds of the heat kernel for all $t > 0$ and almost all $x, y \in M$:

$$\frac{C^{-1}}{V\left(x, t^{1/\beta}\right)} \Phi\left(\frac{d(x,y)}{t^{1/\beta}}\right) \le p_t(x,y) \le \frac{C}{V\left(x, t^{1/\beta}\right)} \widetilde{\Phi}\left(\frac{d(x,y)}{t^{1/\beta}}\right) \tag{6.30}$$

where

$$\widetilde{\Phi}(s) = \begin{cases} \Phi(s), & s \le c^{-1} \\ 0, & s > c^{-1}. \end{cases}$$

Clearly, the functions Φ and $\widetilde{\Phi}$ satisfy the hypotheses of Theorem 4.7. We conclude by Theorem 4.7 that $\beta = \beta^*$ whereas by Lemma 3.1 $\beta^* \ge 2$, which contradicts to $\beta = 1$.

Step 3. Using that $\beta > 1$, let us show that the heat kernel satisfies the following upper bound

$$p_t(x,y) \le \frac{C}{V(x, t^{1/\beta})} \exp\left(-c\left(\frac{d(x,y)}{t^{1/\beta}}\right)^{\beta/(\beta-1)}\right). \tag{6.31}$$

Setting in (6.29) $\lambda = (s/\beta)^{\frac{\beta}{\beta-1}}$ we obtain as in Remark 6.3

$$\Psi(s) = c_\beta s^{\beta/(\beta-1)}$$

so that (6.28) becomes

$$\int_{B(x,r)^c} p_t(x,y)\, d\mu(y) \le C \exp\left(-c\left(\frac{r}{t^{1/\beta}}\right)^{\beta/(\beta-1)}\right). \tag{6.32}$$

On the other hand, by the upper bound in (6.26), we have, for all $t > 0$ and almost all $x, y \in M$,

$$p_t(x,y) \le \frac{C}{V\left(x, t^{1/\beta}\right)}. \tag{6.33}$$

Setting $r = \frac{1}{2}d\,(x,y)$, we obtain from (6.32) and (6.33) that

$$\begin{aligned} p_{2t}(x,y) &= \int_M p_t(x,z)p_t(z,y)\,d\mu(z) \\ &\le \int_{B(x,r)^c} p_t(x,z)p_t(z,y)\,d\mu(z) + \int_{B(y,r)^c} p_t(x,z)p_t(z,y)\,d\mu(z) \qquad (6.34) \\ &\le \frac{C}{V\left(y,t^{1/\beta}\right)} \int_{B(x,r)^c} p_t(x,z)\,d\mu(z) \\ &+ \frac{C}{V\left(x,t^{1/\beta}\right)} \int_{B(y,r)^c} p_t(y,z)\,d\mu(z) \\ &\le \left(\frac{C}{V\left(y,t^{1/\beta}\right)} + \frac{C}{V\left(x,t^{1/\beta}\right)}\right) \exp\left(-c\left(\frac{r}{t^{1/\beta}}\right)^{\beta/(\beta-1)}\right). \qquad (6.35) \end{aligned}$$

By (VD) we have

$$\frac{V\left(x,t^{1/\beta}\right)}{V\left(y,t^{1/\beta}\right)} \le C\left(\frac{d\,(x,y)+t^{1/\beta}}{t^{1/\beta}}\right)^{\alpha} = C\left(1+\frac{r}{t^{1/\beta}}\right)^{\alpha}.$$

Absorbing the polynomial function of $r/t^{1/\beta}$ into the exponential term in (6.35), we obtain (6.31).

Step 4. Now we can prove that $\beta \ge 2$. Indeed, we have the estimate (6.30) where this time

$$\widetilde{\Phi}(s) = \exp\left(-cs^{\beta/(\beta-1)}\right).$$

Since the estimate (6.30) satisfies the hypotheses Theorem 4.7, we obtain $\beta \ge 2$ by the same argument as in Step 2.

Step 5. The lower bound in (6.30) implies that, for all $t > 0$ and almost all $x, y \in M$, such that $d\,(x,y) \le s_0 t^{1/\beta}$,

$$p_t\,(x,y) \ge \frac{c}{V(x,t^{1/\beta})}. \qquad (6.36)$$

Let us show that this implies the following lower bound

$$p_t(x,y) \ge \frac{c}{V(x,t^{1/\beta})} \exp\left(-C\left(\frac{d\,(x,y)}{t^{1/\beta}}\right)^{\beta/(\beta-1)}\right), \qquad (6.37)$$

for all $t > 0$ and almost all $x, y \in M$. Iterating the semigroup identity, we obtain for any positive integer n and real $r > 0$

$$\begin{aligned} p_t(x,y) &= \int_M \dots \int_M p_{\frac{t}{n}}(x,z_1)p_{\frac{t}{n}}(z_1,z_2)\dots p_{\frac{t}{n}}(z_{n-1},y)d\mu(z_{n-1})\dots d\mu(z_1) \\ &\ge \int\limits_{B(x_1,r)} \dots \int\limits_{B(x_{n-1},r)} p_{\frac{t}{n}}(x,z_1)p_{\frac{t}{n}}(z_1,z_2)\dots p_{\frac{t}{n}}(z_{n-1},y)d\mu(z_{n-1})\dots d\mu(z_1), \end{aligned} \qquad (6.38)$$

where $\{x_i\}_{i=0}^n$ is a chain connecting x and y that satisfies (6.25) (see Fig. 8).

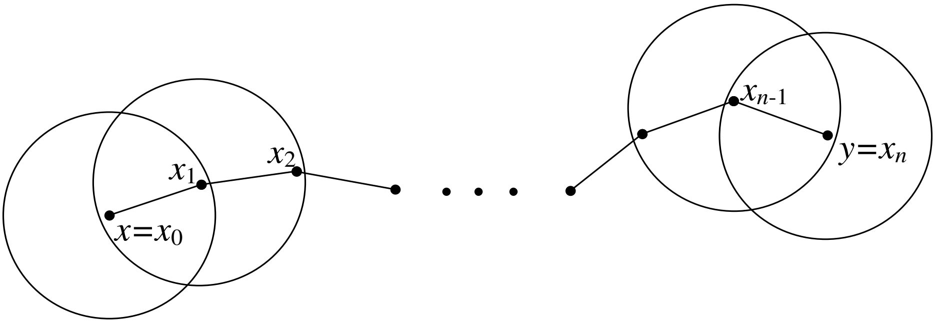

FIGURE 8. Chain $\{x_i\}$

Denote for simplicity $z_0 = x$ and $z_n = y$. Setting

$$r = \frac{d(x,y)}{n} \tag{6.39}$$

and noticing that $z_i \in B(x_i, r), 0 \le i \le n-1$, we obtain by the triangle inequality and (6.25)

$$d(z_i, z_{i+1}) \le d(x_i, x_{i+1}) + 2r \le C' \frac{d(x,y)}{n}$$

where $C' = C + 2$. Next, we would like to use (6.36) to estimate $p_{t/n}(z_i, z_{i+1})$ from below. For that, the following condition must be satisfied:

$$d(z_i, z_{i+1}) \le s_0 \left(\frac{t}{n}\right)^{1/\beta},$$

which will follow from

$$C' \frac{d(x,y)}{n} \le s_0 \left(\frac{t}{n}\right)^{1/\beta}.$$

Absorbing the constants C' and s_0 into one, we see that the latter condition is equivalent to

$$n \ge c \left(\frac{d(x,y)}{t^{1/\beta}}\right)^{\frac{\beta}{\beta-1}}. \tag{6.40}$$

If $d(x,y) \le s_0 t^{1/\beta}$ then (6.37) follows immediately from (6.36). Assume in the sequel that $d(x,y) > s_0 t^{1/\beta}$ and choose n to be the least positive integer satisfying (6.40), that is

$$n \asymp \left(\frac{d(x,y)}{t^{1/\beta}}\right)^{\frac{\beta}{\beta-1}} \tag{6.41}$$

This and (6.39) clearly imply that

$$r \asymp \left(\frac{t}{n}\right)^{1/\beta}. \tag{6.42}$$

Then we have by (6.36) and (VD)

$$p_{\frac{t}{n}}(z_i, z_{i+1}) \geq \frac{c}{V\left(z_i, (t/n)^{1/\beta}\right)} \geq \frac{c}{V(z_i, r)}. \tag{6.43}$$

Since by (VD)

$$\frac{V(z_i, r)}{V(x_i, r)} \leq C\left(\frac{d(z_i, x_i) + r}{r}\right)^{\alpha} \leq C2^{\alpha},$$

it follows from (6.43) that

$$p_{\frac{t}{n}}(z_i, z_{i+1}) \geq \frac{c}{V(x_i, r)}.$$

Using (6.38), (6.42), (VD), (RVD), and (6.41), we obtain

$$\begin{aligned} p_t(x,y) &\geq \int\limits_{B(x_1,r)} \cdots \int\limits_{B(x_{n-1},r)} \frac{c^n d\mu(z_{n-1}) \ldots d\mu(z_1)}{V(x,r)\, V(x_1,r) \ldots V(x_{n-1},r)} \\ &= \frac{c^n}{V(x,r)} \geq c' \frac{c^n}{V\left(x, (t/n)^{1/\beta}\right)} \\ &= \frac{c'}{V\left(x, t^{1/\beta}\right)} \frac{c^n V\left(x, t^{1/\beta}\right)}{V\left(x, (t/n)^{1/\beta}\right)} \geq c' \frac{c^n n^{\alpha'/\beta}}{V\left(x, t^{1/\beta}\right)} \\ &\geq \frac{c'}{V\left(x, t^{1/\beta}\right)} \exp(-Cn) \\ &\geq \frac{c'}{V\left(x, t^{1/\beta}\right)} \exp\left(-C\left(\frac{d(x,y)^{\beta}}{t}\right)^{\frac{1}{\beta-1}}\right). \end{aligned}$$

Comparing (6.26) with (6.31) and (6.37), we obtain (6.27). □

Corollary 6.6. *Under the hypotheses of Theorem* 6.5, *we have* $\mathcal{E}(u) \asymp E_{\beta}(u)$ *for all* $u \in L^2(M)$. *Consequently,* $\mathcal{F} = W^{\beta/2,2}$.

Proof. Indeed, by Lemma 4.2 we have $\mathcal{E}(u) \geq cE_{\beta}(u)$. Using the upper bound (6.31) and Lemma (4.5), we obtain $\mathcal{E}(u) \leq CE_{\beta}(u)$, which finishes the proof. □

References

[1] D.G. Aronson, Non-negative solutions of linear parabolic equations Ann. Scuola Norm. Sup. Pisa. Cl. Sci. (4) **22** (1968), 607–694. Addendum **25** (1971), 221–228.

[2] M.T. Barlow, *Diffusions on fractals*, Lect. Notes Math. **1690**, Springer, 1998, 1–121.

[3] M.T. Barlow and R.F. Bass, Brownian motion and harmonic analysis on Sierpínski carpets, Canad. J. Math. (4) **51** (1999), 673–744.

[4] M.T. Barlow, R.F. Bass, Z.-Q. Chen and M. Kassmann, Non-local Dirichlet forms and symmetric jump processes, Trans. Amer. Math. Soc. **361** (2009), 1963–1999.

[5] M. Barlow, T. Coulhon and T. Kumagai, Characterization of sub-Gaussian heat kernel estimates on strongly recurrent graphs, Comm. Pure Appl. Math., **58** (2005), 1642–1677.

[6] M.T. Barlow and E.A. Perkins, Brownian motion on the Sierpínski gasket, Probab. Theory. Related Fields **79** (1988), 543–623.

[7] A. Bendikov and L. Saloff-Coste, On-and-off diagonal heat kernel behaviors on certain infinite-dimensional local Dirichlet spaces, Amer. J. Math. **122** (2000), 1205–1263.

[8] G. Carron, Inégalités isopérimétriques de Faber-Krahn et conséquences, in: *Actes de la table ronde de géométrie différentielle* (*Luminy,* 1992) Collection SMF Séminaires et Congrès, **1** (1996), 205–232.

[9] E.A. Carlen and S. Kusuoka and D.W. Stroock, Upper bounds for symmetric Markov transition functions, Ann. Inst. Henri. Poincaré Probab. Statist. **23**(1987), 245–287.

[10] I. Chavel, *Eigenvalues in Riemannian geometry*, Academic Press, New York, 1984.

[11] T. Coulhon, Ultracontractivity and Nash type inequalities, J. Funct. Anal. **141** (1996), 510–539.

[12] T. Coulhon, Heat kernel and isoperimetry on non-compact Riemannian manifolds, in: Heat kernels and analysis on manifolds, graphs, and metric spaces, 65–99, Contemp. Math., **338** Amer. Math. Soc., Providence, RI, 2003.

[13] M. Cowling and S. Meda, Harmonic analysis and ultracontractivity. Trans. Amer. Math. Soc. **340** (1993), 733–752.

[14] E.B. Davies, *Heat kernels and spectral theory*, Cambridge University Press, 1989.

[15] M. Fukushima and Y. Oshima Y. and M. Takeda, *Dirichlet forms and symmetric Markov processes*, De Gruyter, Studies in Mathematics, 1994.

[16] M. Fukushima and T. Shima, On a spectral analysis for the Sierpiński gasket. Potential Anal. **1** (1992), 1–35.

[17] A. Grigor'yan, Heat kernel upper bounds on a complete non-compact manifold, Rev. Mat. Iberoamericana **10**(1994), 395–452.

[18] A. Grigor'yan, Estimates of heat kernels on Riemannian manifolds, London Math. Soc. Lecture Note Ser., **273** (1999),140–225.

[19] A. Grigor'yan, Heat kernels on metric measure spaces, to appear in "Handbook of Geometric Analysis No.2" (ed. L. Ji, P. Li, R. Schoen, L. Simon), Advanced Lectures in Math., IP.

[20] A. Grigor'yan and J. Hu, Off-diagonal upper estimates for the heat kernel of the Dirichlet forms on metric spaces, Invent. Math. **174** 2008, 81–126.

[21] A. Grigor'yan and J. Hu, Upper bounds of heat kernels on doubling spaces, preprint 2008.

[22] A. Grigor'yan, J. Hu and K.-S. Lau, Heat kernels on metric-measure spaces and an application to semilinear elliptic equations, Trans. Amer. Math. Soc. **355** (2003), 2065–2095.

[23] A. Grigor'yan, J. Hu and K.-S. Lau, Equivalence conditions for on-diagonal upper bounds of heat kernels on self-similar spaces, J. Func. Anal. **237** (2006), 427–445.

[24] A. Grigor'yan and T. Kumagai, On the dichotomy in the heat kernel two sided estimates. In: Analysis on Graphs and its Applications (P. Exner et al. (eds.)), Proc. of Symposia in Pure Math. **77**, pp. 199–210, Amer. Math. Soc. 2008.

[25] A. Grigor'yan and A. Telcs, Sub-Gaussian estimates of heat kernels on infinite graphs, Duke Math. J. **109**(2001), 451–510.

[26] B.M. Hambly and T. Kumagai, Transition density estimates for diffusion processes on post critically finite self-similar fractals, Proc. London Math. Soc. (3)**79** (1999), 431–458.

[27] W. Hebisch and L. Saloff-Coste, On the relation between elliptic and parabolic Harnack inequalities, Ann. Inst. Fourier (Grenoble) **51** (2001), 1437–1481.

[28] A. Jonsson, Brownian motion on fractals and function spaces, Math. Zeit. **222** (1996), 495–504.

[29] J. Jorgenson and W. Lynne, ed. *The ubiquitous heat kernel.* Contemporary Math. **398**, AMS, 2006.

[30] J. Kigami, *Analysis on fractals*, Cambridge University Press, Cambridge, 2001.

[31] J. Kigami, Local Nash inequality and inhomogeneous of heat kernels, Proc. London Math. Soc. **89** (2004), 525–544.

[32] P.Li and S.T. Yau, On the parabolic kernel of the Schrödinger operator, Acta Math. **156**(1986), 153–201.

[33] J. Nash, Continuity of solutions of parabolic and elliptic equations, Amer. J. Math. **80** (1958), 931–954.

[34] K. Pietruska-Pałuba, On function spaces related to fractional diffusion on d-sets, Stochastics and Stochastics Reports, **70** (2000), 153–164.

[35] F.O. Porper and S.D. Eidel'man, Two-side estimates of fundamental solutions of second-order parabolic equations and some applications, Russian Math. Surveys, **39** (1984), 119–178.

[36] L. Saloff-Coste, *Aspects of Sobolev-type inequalities*, Cambridge University Press, Cambridge, 2002.

[37] L. Saloff-Coste, A note on Poincaré, Sobolev, and Harnack inequalities, Internat. Math. Res. Notices 1992, no. 2, 27–38.

[38] R. Schoen and S.-T Yau, *Lectures on Differential Geometry*, International Press, 1994.

[39] D.W. Stroock, Estimates on the heat kernel for the second-order divergence form operators, in: *Probability theory. Proceedings of the* 1989 *Singapore Probability Conference held at the National University of Singapore, June* 8–16 1989, ed. L.H.Y. Chen, K.P. Choi, K. Hu and J.H. Lou, Walter De Gruyter, 1992, 29–44.

[40] M. Tomisaki, Comparison theorems on Dirichlet norms and their applications, Forum Math. **2** (1990), 277–295.

[41] N.Th. Varopoulos, Hardy-Littlewood theory for semigroups, J. Funct. Anal. **63** (1985), 240–260.

[42] N.Th. Varopoulos and L. Saloff-Coste and T. Coulhon, *Analysis and geometry on groups*, Cambridge Tracts in Mathematics, **100**. Cambridge University Press, Cambridge, 1992.

[43] F.-Y. Wang, Functional inequalities for empty essential spectrum, J. Funct. Anal. **170** (2000), 219–245.

Alexander Grigor'yan
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
D-33501 Bielefeld, Germany
e-mail: `grigor@math.uni-bielefeld.de`

Jiaxin Hu
Department of Mathematical Sciences
Tsinghua University
Beijing 100084, China
e-mail: `hujiaxin@mail.tsinghua.edu.cn`

Ka-Sing Lau
Department of Mathematics
the Chinese University of Hong Kong
Shatin, N.T., Hong Kong
e-mail: `kslau@math.cuhk.edu.hk`

Progress in Probability, Vol. 61, 45–70

Self-similarity and Random Walks

Vadim A. Kaimanovich

Abstract. This is an introductory level survey of some topics from a new branch of fractal analysis – the theory of self-similar groups. We discuss recent works on random walks on self-similar groups and their applications to the problem of amenability for these groups.

Mathematics Subject Classification (2000). 20E22, 43A07, 60J05.

Keywords. Self-similar group, random walk, amenability, entropy.

Introduction

The purpose of this paper is to give a brief survey of some ideas and methods associated with new progress in understanding the so-called *self-similar groups*. This class of groups consists of automorphisms of homogeneous rooted trees defined in a simple recursive way. A good account of the initial period of the theory can be found in the survey [BGN03] by Bartholdi, Grigorchuk and Nekrashevych and in the recent monograph of Nekrashevych [Nek05] (their authors are among the most active contributors to this field).

Self-similar groups have a natural interpretation in terms of fractal geometry; their *limit sets* are very interesting fractal sets (see the recent papers [NT08, RT09] for a study of Laplacians on limit sets). These groups also arise as *iterated monodromy groups* of rational endomorphisms of the Riemann sphere, which, for instance, led to a recent solution of an old problem from rational dynamics [BN06].

Self-similar groups are often quite unusual from the point of view of the traditional group theory. This has both advantages and disadvantages. On one hand, by using self-similar groups it is easy to construct examples which may otherwise be much less accessible (for instance, the famous *Grigorchuk group of intermediate growth* [Gri80, Gri85, dlH00] has a very simple self-similar presentation). On the other hand, even the simplest group theoretical questions for self-similar groups may be quite hard. Already finding a self-similar realization of a free group is very far from being obvious [BS98, GM05, VV07].

Another question of this kind is whether a given self-similar group is *amenable* (amenability introduced by von Neumann [vN29] is, in a sense, the most natural generalization of finiteness, and it plays a fundamental role in group theory). There are numerous characterizations of amenability in various terms. In particular, it is known to be equivalent to existence of a *random walk* on the group with *trivial behaviour at infinity* (≡ *trivial Poisson boundary*) [Fur73, Ros81, KV83]. It turns out that self-similar groups may have random walks which are also self-similar in a certain sense, and it is this self-similarity that can be used in order to prove triviality of the Poisson boundary, and therefore establish amenability of the underlying group.

This idea was first used by Bartholdi and Viràg [BV05] for proving amenability of the so-called *Basilica group* $\mathcal{B}$. This group first studied by Grigorchuk and Żuk [GŻ02a] has a very simple matrix presentation and also arises as the iterated monodromy group of the map $z \mapsto z^2 - 1$ (known as the Basilica map, whence the name). In particular, Grigorchuk and Żuk proved that $\mathcal{B}$ is not *subexponentially elementary*, which made especially interesting the question about its amenability. The approach of Bartholdi and Viràg was further developed by Kaimanovich [Kai05] who used the *entropy theory* of random walks (which provides a simple criterion of triviality of the Poisson boundary) in combination with a contraction property for the asymptotic entropy of random walks on self-similar groups (the *"Münchhausen trick"*).

An important ingredient of this technique is a link (established in [Kai05]) between self-similarity and the so-called *random walks with internal degrees of freedom (RWIDF)* [KS83] also known under the names of *matrix-valued random walks* [CW89] or of *covering Markov chains* [Kai95]. These are group invariant Markov chains which take place on the product of a group by a certain parameter set. Any random walk on a self-similar group naturally gives rise to a random walk with internal degrees of freedom parameterized by the alphabet of the action tree of the group. In turn, this RWIDF, when restricted to the copy of the group corresponding to a fixed value of the freedom parameter, produces a new random walk on the self-similar group (as it is pointed out in [GN07], this transformation corresponds to the classical operation of taking the *Schur complement* of a matrix). It is the interplay between the original and the new random walks, which allows one to apply the Münchhausen trick.

This technique was recently applied by Bartholdi, Kaimanovich and Nekrashevych [BKN08] to prove amenability of all *self-similar groups generated by bounded automata* (this class, in particular, contains the Basilica group).

Added in proof. The recent preprint "Amenability of linear-activity automaton groups" by Gideon Amir, Omer Angel and Bálint Virág (arXiv:0905.2007) contains a proof of amenability of all self-similar groups generated by automata of linear growth. It is based on an extension of the methods from the above paper.

In this survey we attempt to give a historical and conceptual overview of these developments without going into technical details, so that it should hopefully be suitable for a first acquaintance with the subject. Structurally, our presentation is split into three parts.

In Section 1 we discuss the notion of self-similarity, introduce self-similar groups in general and the subclass of self-similar groups generated by bounded automata.

Further in Section 2 we briefly discuss the notion of amenability of a group.

Finally, in Section 3 we analyze random walks on self-similar groups, and show how they can be used for establishing amenability of self-similar groups.

The presentation is based on a talk given at the "Fractal Geometry and Stochastics IV" conference (and on several other occasions as well). I would like to thank the organizers of this meeting for a very interesting and inspiring week in Greifswald.

1. Self-similar groups

1.A. And so ad infinitum...

The modern idea of self-similarity is best described by the following quote from *On Poetry: a Rhapsody* by Jonathan Swift (1733)[1]:

> So, naturalists observe, a flea
> Has smaller fleas that on him prey;
> And these have smaller still to bite 'em,
> And so proceed *ad infinitum.*

On a more formal level, the simplest self-similarity assumptions are:

- a part is similar to the whole;
- the whole is a union of such parts;
- these parts are pairwise disjoint.

These assumptions naturally lead to the most fundamental self-similar structure, that of a *rooted homogeneous tree.* Such trees are, for instance, skeletons of iterated function systems satisfying the open set condition (e.g., see [Fal03]); the classical Cantor set is produced by such a system. More precisely, let $X \cong \{1, 2, \dots, d\}$ be a finite set called the *alphabet.* Denote by X^* the set of all finite words in the alphabet X (including the empty word $\varnothing$). In other terms, X^* is the *free monoid* generated by the set X (the composition being the concatenation $(w, w') \mapsto ww'$). The associated rooted homogeneous tree $T = T(X)$ is the (right) Cayley graph of the free monoid X^* (so that one connects w to wx by an edge for all $w \in X^*, x \in X$). The tree $T(X) \cong X^*$ is split into *levels* $T_n \cong X^n$ (the set of

[1] It was *naturally extended* by Augustus de Morgan in his *Budget of Paradoxes* (1872):

> Great fleas have little fleas upon their backs to bite 'em,
> And little fleas have lesser fleas, and so *ad infinitum.*
> And the great fleas themselves, in turn have, greater fleas to go on;
> While these again have greater still, and greater still, and so on.

This is a description of what is nowadays called the *natural extension* of a non-invertible dynamical system. See [Kai03] for its applications in the context of fractal sets.

words of length n). The level T_0 consists only of the empty word $\varnothing$, which is the root of T. Each vertex $w \in T \cong X^*$ is the root of the subtree T_w which consists of all the words beginning with w. The map $w' \mapsto ww'$ provides then a canonical identification of the trees T and T_w, see Figure 1, where $X = \{a, b\}$.

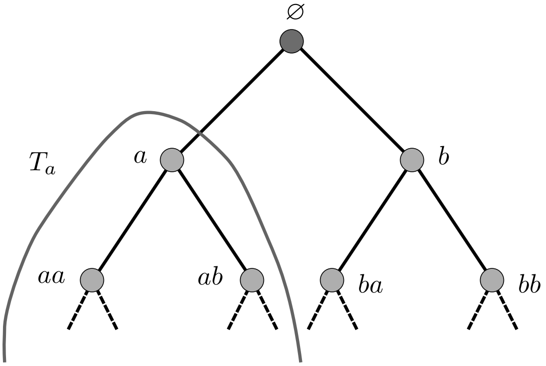

Figure 1

1.B. Generalized permutation matrices

The group $\mathfrak{G} = \mathsf{Aut}(T)$ of automorphisms of the tree T obviously preserves each level of T. In particular, it acts by permutations on the first level $X \cong T_1$, i.e., there is a homomorphism $g \mapsto \sigma = \sigma^g$ from $\mathfrak{G}$ to $\mathsf{Sym}(X)$ (the permutation group on X). Another piece of data associated with any automorphism $g \in \mathfrak{G}$ is a collection of automorphisms $\{g_x\}_{x \in X}$ indexed by the alphabet X. Indeed, if $y = \sigma^g(x)$, then g establishes a one-to-one correspondence between the corresponding subtrees T_x and T_y rooted at the points x and y, respectively. Since both subtrees T_x and T_y are canonically isomorphic to the full tree T, the map $g : T_x \to T_y$ is conjugate to an automorphism of T denoted by g_x (in terms of Swift's description above, g_x describes what happens on the back of a first-order flea when it moves from position x to position $y = \sigma^g(g)$), see Figure 2.

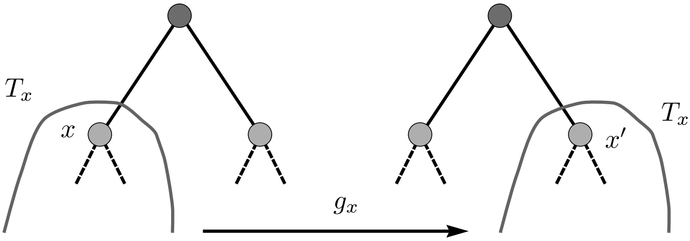

Figure 2

Conversely, a permutation $\sigma \in \mathsf{Sym}(X)$ and a collection $\{g_x\}_{x\in X}$ of elements of $\mathfrak{G}$ uniquely determine the associated automorphism $g \in \mathfrak{G}$. In algebraic terms it means that $g \mapsto \big(\sigma^g; \{g_x\}_{x\in X}\big)$ is an isomorphism of the group $\mathfrak{G}$ and the *semi-direct product* $\mathsf{Sym}(X) \leftthreetimes \mathfrak{G}^X$ (in yet another terminology, $\mathfrak{G}$ is isomorphic to the *permutational wreath product* $\mathfrak{G} \wr \mathsf{Sym}(X)$). There is a very convenient way of visualizing this structure by means of *generalized permutation matrices.*

Recall that the usual *permutation matrix* M^σ associated with a permutation $\sigma \in \mathsf{Sym}(X)$ is a $|X| \times |X|$ matrix with entries

$$M^\sigma_{xy} = \begin{cases} 1\,, & \text{if } y = \sigma(x), \\ 0\,, & \text{otherwise} \end{cases} ,$$

and that the map $\sigma \mapsto M^\sigma$ is a group isomorphism. In the same way we shall present the data $\big(\sigma^g; \{g_x\}_{x\in X}\big)$ by the *generalized permutation matrix* M^g with entries

$$M^g_{xy} = \begin{cases} g_x\,, & \text{if } y = \sigma^g(x), \\ 0\,, & \text{otherwise.} \end{cases}$$

For instance, the automorphism g described in Figure 3 is presented by the matrix

$$M^g = \begin{pmatrix} 0 & g_1 \\ g_2 & 0 \end{pmatrix} .$$

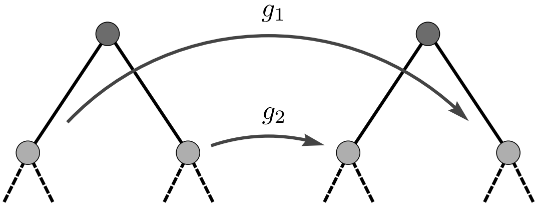

Figure 3

More generally, given an arbitrary group G, we shall denote by

$$\mathsf{Sym}(X; G) = G \wr \mathsf{Sym}(X) = \mathsf{Sym}(X) \leftthreetimes G^X$$

the *group of generalized permutation matrices* of order $|X|$ with non-zero entries from the group G. The group operation here is the usual matrix multiplication, the only difference with ordinary permutation matrices being that the matrix elements are multiplied according to the group law of G. Obviously, application of the *augmentation map* (which consists in replacing all group elements with 1) to a generalized permutation matrix yields a usual permutation matrix, which corresponds to the natural homomorphism of $\mathsf{Sym}(X; G)$ onto $\mathsf{Sym}(X)$. We can now sum up the above discussion by saying that *there is a natural isomorphism of the group* $\mathfrak{G} = \mathsf{Aut}(T)$ *of automorphisms of the tree* $T = T(X)$ *and of the generalized permutation group* $\mathsf{Sym}(X; \mathfrak{G})$. It is this isomorphism that embodies the self-similarity properties of the group $\mathfrak{G}$.

1.C. Self-similar groups and matrix presentations

Definition. A countable subgroup $G \subset \mathfrak{G}$ is *self-similar* if the restriction of the isomorphism $\mathfrak{G} \to \mathsf{Sym}(X;\mathfrak{G})$ to G induces an embedding $G \hookrightarrow \mathsf{Sym}(X;G)$; in other words, if all entries of the matrices M^g, $g \in G$ belong to G. Note that, rigorously speaking, self-similarity is a property of the *embedding* $G \subset \mathfrak{G}$ rather than of the group G only. The embedding $G \hookrightarrow \mathsf{Sym}(X;G)$ need *not* be surjective (see the example below).

Example 1. The *adding machine* (isomorphic to the group $\mathbb{Z}$) is generated by the transformation $a : z \mapsto z+1$ on the ring $\mathbb{Z}_2$ of 2-*adic integers* $\varepsilon_0 + \varepsilon_1 \cdot 2 + \cdots + \varepsilon_n \cdot 2^n + \cdots$, where the digits ε_i take values 0 or 1. Depending on the values of initial digits, it acts in the following way:

$$\begin{array}{cccccccc} 0 & \varepsilon_1 & \varepsilon_2 & \varepsilon_3\dots & \mapsto & 1\ \varepsilon_1\ \varepsilon_2\ \varepsilon_3\dots\ , \\ 1 & 0 & \varepsilon_2 & \varepsilon_3\dots & \mapsto & 0\ 1\ \varepsilon_2\ \varepsilon_3\dots\ , \\ 1 & 1 & 0 & \varepsilon_3\dots & \mapsto & 0\ 0\ 1\ \varepsilon_3\dots\ , \\ & & & & \dots\dots & \end{array}$$

We can think of sequences $(\varepsilon_0, \varepsilon_1, \dots)$ as of boundary points of the binary rooted tree $T = T(X)$ of the alphabet $X = \{0,1\}$. The transformation a extends to an automorphism of T, and, as one can easily see from its symbolic description above, the associated generalized permutation matrix is

$$M^a = \begin{pmatrix} 0 & 1 \\ a & 0 \end{pmatrix} .$$

Thus, the infinite cyclic group $\langle a \rangle \cong \mathbb{Z}$ generated by the transformation a is self-similar (as a subgroup of the full group of automorphisms $\mathfrak{G}$).

Note that the automorphism a is completely determined by the matrix M^a. Indeed, the augmentation map applied to M^a produces the usual permutation matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ which describes the permutation by which a acts on the first level of the tree T. Further, by substituting M^a for a and the identity matrix for 1 in M^a one obtains the matrix

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ a & 0 & 0 & 0 \end{pmatrix} ,$$

augmentation of which produces the order 4 permutation matrix

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

describing the action of a on the second level of T. By recursively repeating this procedure one obtains the action of a on all levels of T, i.e., an automorphism of the whole tree T.

The example above suggests the following way of defining self-similar groups by their matrix presentations. Fix a finite (or countable, for infinitely generated groups) set K (the future set of generators of a self-similar group), and assign to each $\kappa \in K$ a generalized permutation matrix M^κ whose non-zero entries are words in the alphabet consisting of letters from K and their inverses (the entry associated with the empty word is 1). By replacing the non-zero entries of matrices M^κ with corresponding products of the associated matrices and their inverses we obtain generalized permutation matrices of order $|X|^2$, etc. The usual permutation matrices obtained from them by the augmentation map determine then the action of elements from K on all levels of the tree T, i.e., the corresponding automorphisms of T. See [BG00] or [Nek05] for more on recursions of this kind.

A particular case of this construction arises in the situation when all non-zero entries of matrices M^κ are elements of the set K. In this case the assignment $\kappa \mapsto M^\kappa$ amounts to a map $(\kappa, x) \mapsto (\lambda, y)$ of the product $K \times X$ to itself, i.e., to an *automaton.* Here $y = \sigma(x)$ for the permutation $\sigma = \sigma(M^\kappa)$ determined by the matrix M^κ, and λ is the matrix entry M^κ_{xy}. The self-similar group obtained in this way is called an *automaton group.*[2]

1.D. The Basilica group

The *Basilica group* $\mathcal{B}$ is determined by the matrix presentation

$$a \mapsto \begin{pmatrix} b & 0 \\ 0 & 1 \end{pmatrix}, \qquad b \mapsto \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix}.$$

The aforementioned recursion for this group looks in the following way:

$$a \mapsto \begin{pmatrix} b & 0 \\ 0 & 1 \end{pmatrix} \mapsto \begin{pmatrix} 0 & a & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \mapsto \cdots,$$

$$b \mapsto \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix} \mapsto \begin{pmatrix} 0 & 0 & b & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \mapsto \cdots,$$

and the associated automaton is

$$\begin{cases} (a,1) & \mapsto & (b,1) \\ (a,2) & \mapsto & (e,2) \\ (b,1) & \mapsto & (a,2) \\ (b,2) & \mapsto & (e,1) \\ (e,1) & \mapsto & (e,1) \\ (e,2) & \mapsto & (e,2) . \end{cases}$$

[2] Usually, when talking about automata groups, one tacitly assumes that the corresponding automaton is finite, i.e., the generating set K is finite.

Here $\{1,2\} = X$ is the alphabet of the binary tree T, and $K = \{a, b, e\}$, where e is the group identity determined by the substitution $e \mapsto \begin{pmatrix} e & 0 \\ 0 & e \end{pmatrix}$.

The group $\mathcal{B}$ was first studied by Grigorchuk and Żuk [GŻ02a] (see below for its algebraic properties). The name *Basilica* comes from the fact that it also appears as the *iterated monodromy group* of the rational map $z \mapsto z^2 - 1$ on the Riemann sphere $\overline{\mathbb{C}}$.

This latter notion was introduced by Nekrashevych who created a very fruitful link between the theory of self-similar groups and *rational dynamics.* Namely, given a rational map $\varphi : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ of degree d, a generic point $z \in \overline{\mathbb{C}}$ has precisely d preimages, each of which also has d preimages, etc. Thus, attached to a generic point $z \in \overline{\mathbb{C}}$ is the rooted homogeneous tree T_z of its preimages. One can move the preimage tree along any continuous curve consisting of generic points. However, if z follows a non-contractible loop, it may happen that, although the preimage tree T_z returns to its original position, it undergoes a certain non-trivial *monodromy transformation.* Thus, there is a homomorphism of the fundamental group of the connected component of z in the set of generic points to the group of automorphisms of the preimage tree T_z. The resulting subgroup of $\mathsf{Aut}(T_z)$ is called the *iterated monodromy group* of the map φ, see [Nek05] for more details.

Now, in rational dynamics the map $z \mapsto z^2 - 1$ is called the *Basilica map* [Bie90], because its *Julia set* (a subset of the Riemann sphere which, in a sense, consists of the limit points of this map) looks similar to *Basilica di San Marco* in Venice (together with its reflection in the water), see Figure 4. This Julia set also arises as the limit set of the Basilica group $\mathcal{B}$.

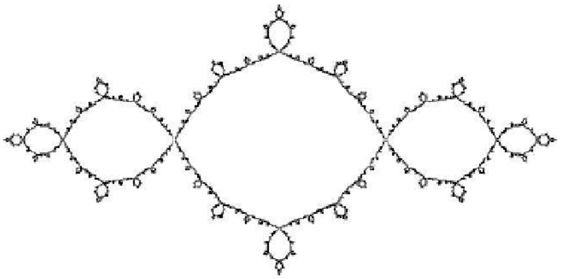

Figure 4

Interesting algebraic and analytic properties of the Basilica group obtained in [GŻ02a, GŻ02b, BG02] made especially relevant the question about its *amenability* formulated in [GŻ02a], also see [BGN03].

1.E. Bounded automatic automorphisms and Mother groups

The Basilica group $\mathcal{B}$ actually belongs to a certain natural subclass of the class of self-similar groups which we shall now describe.

As we have seen in Section 1.B, given an automorphism $g \in \mathfrak{G} = \mathsf{Aut}(T)$, any symbol $x \in X$ determines an associated automorphism $g_x \in \mathfrak{G}$. In the same way such an automorphism $g_w \in \mathfrak{G}$ (the *state* of g at the point w) can be defined for an arbitrary word $w \in T \cong X^*$, by restricting the automorphism g to the subtree T_w with the subsequent identification of both T_w and its image $g(T_w) = T_{g(w)}$ with T. Equivalently, g_w can be obtained by recursively applying the presentation $g \mapsto M^g$, namely, g_w is the non-zero entry of the row R_w of the $|X|^{|w|} \times |X|^{|w|}$ matrix obtained by recursive expansion $g \mapsto M^g \mapsto \cdots$.

If the *set of states* of g

$$\mathsf{S}(g) = \{g_w : w \in T\} \subset \mathfrak{G}$$

is finite, then the automorphism g is called *automatic.* The set of all automatic automorphisms of the tree T forms a countable subgroup $\mathfrak{A} = \mathfrak{A}(X)$ of $\mathfrak{G} = \mathfrak{G}(X)$.

An automorphism g is called *bounded* if the sets $\{w \in X^n : g_w \neq e\}$ have uniformly bounded cardinalities over all n. The set of all bounded automorphisms also forms a subgroup $\mathfrak{B} = \mathfrak{B}(X)$ of $\mathfrak{G} = \mathfrak{G}(X)$. We denote by $\mathfrak{BA} = \mathfrak{BA}(X) = \mathfrak{B}(X) \cap \mathfrak{A}(X)$ the *group of all bounded automatic automorphisms* of the homogeneous rooted tree T.

It is easy to see that the generators a, b of the Basilica group $\mathcal{B}$ described in Section 1.D are both automatic and bounded in the above sense, so that $\mathcal{B} \subset \mathfrak{BA}$ (here and on several occasions below we omit the alphabet X from our notation). More generally, any group generated by a (finite) automaton (see Section 1.C) is a subgroup of $\mathfrak{A}$. An automaton is called *bounded* if its group is contained in $\mathfrak{B}$ (therefore, in $\mathfrak{BA}$). The class of *groups generated by bounded automata* was defined by Sidki in [Sid00]. Obviously, all these groups are subgroups of $\mathfrak{BA}$. Most of the well-studied groups of finite automata belong to this class (see [BKN08] for examples).

Groups generated by bounded automata also appear naturally in connection with fractal geometry. It was proved in [BN03] that every such group is *contracting*, and a contracting group is generated by bounded automata if and only if the boundary of its *tile* is finite. This technical condition implies that the *limit space* (see [Nek05]) of such a group belongs to the well-studied class of *nested fractals* (see [Kig01]). This is the class of fractals on which the Brownian motion is best understood. This shows an interesting connection between the most well-studied class of self-similar groups and the class of fractals with most well-understood analysis (see [NT08] for more details).

It turns out that the class of groups generated by bounded automata contains a countable family of groups which have certain *universality properties* with respect to this class.

Let X be a finite set with a distinguished element $o \in X$, and put $\overline{X} = X \setminus \{o\}$. Set $A = \mathsf{Sym}(X)$ and $B = \mathsf{Sym}(\overline{X}; A)$, and recursively embed the groups A and B into $\mathfrak{G}(X)$ by the matrix presentations

$$M^a = \phi_A(a) \ , \quad M^b = \begin{pmatrix} b & 0 \\ 0 & \phi_B(b) \end{pmatrix} \ ,$$

where $\phi_A(a), \phi_B(b)$ are, respectively, the permutation and the generalized permutation matrices corresponding to $a \in A, b \in B$. Then the *Mother group* $\mathfrak{M} = \mathfrak{M}(X) = \langle A, B \rangle$ is the subgroup of $\mathfrak{G} = \mathfrak{G}(X)$ generated by the finite groups A and B.

A direct verification shows that both groups A, B are contained in $\mathfrak{BA}$, whence *the group* $\mathfrak{M}(X)$ *is a subgroup of* $\mathfrak{BA}(X)$. On the other hand, as it was proved in [BKN08], *any finitely generated subgroup of* $\mathfrak{BA}(X)$ *can be embedded as a subgroup into the generalized permutation group* $\mathsf{Sym}(X^N; \mathfrak{M}(X^N))$ *for some integer* N.

Thus, in view of the fact that amenability is preserved by *elementary operations* (see Section 2.C below), amenability of the groups $\mathfrak{BA}(X)$ for all finite sets X (therefore, amenability of all groups generated by bounded automata) would follow from amenability just of all the Mother groups $\mathfrak{M}(X)$.

It is worth noting that the groups generated by bounded automata form a subclass of the class of *contracting self-similar groups* (see [BN03, Nek05]). It is still an open question whether all contracting groups are amenable. However, Nekrashevych [Nek08] recently established a weaker property: *contracting groups contain no free groups with* ≥ 2 *generators.*

2. Amenability

2.A. From finite to infinite: von Neumann, Day and Reiter

Finite groups can be characterized as those discrete groups which have a *finite invariant measure.* In other words, a discrete group G is finite if and only if the natural action of G by translations on the space $\ell^1_{+,1}(G)$ of positive normalized elements from $\ell^1(G)$ has a fixed point. This trivial observation suggests two ways of "extending" the finiteness property to infinite groups. One can look either for fixed points in a bigger space, or for approximative invariance instead of exact one.

The first idea was implemented by John von Neumann [vN29], according to whose definition *amenable groups are those which admit a translation invariant mean*[3]. A *mean* on G is a *finitely additive* "probability measure", in other words, an element of the space $[\ell^\infty]^*_{+,1}$ of positive normalized functionals on $\ell^\infty(G)$. Usual measures on G are also means, but if G is infinite, then there are many more means than measures (which corresponds to the fact that in the infinite case ℓ^1 is significantly "smaller" than its second dual space $[\ell^1]^{**} = [\ell^\infty]^*$).

[3] Actually, the original term used by von Neumann was the German *meßbare Gruppe*, which means "measurable group" in English. It was later replaced in German with *mittelbare* (cf.

Means being highly non-constructive objects, the other way was explored (surprisingly, much later than the original definition of von Neumann) by Reiter [Rei65] who introduced (under the name P_1) what is nowadays known as *Reiter's condition* for a group G: *there exists an approximatively invariant sequence of probability measures on G,* in other words, *there exists a sequence of probability measures λ_n on G such that*

$$\|\lambda_n - g\lambda_n\| \underset{n\to\infty}{\longrightarrow} 0 \qquad \forall\, g \in G ,$$

where $\|\cdot\|$ denotes the total variation norm. He proved that the above condition is in fact equivalent to amenability as defined by von Neumann.

Example 2. The sequence of *Cesaro averaging measures*

$$\lambda_n = \frac{1}{n+1}\big(\delta_0 + \delta_1 + \cdots + \delta_n\big) ,$$

on the group of integers $\mathbb{Z}$ is approximatively invariant (here δ_n denotes the unit mass at the point n). Thus, $\mathbb{Z}$ is amenable.

2.B. Other definitions

There is a lot of other (equivalent) definitions of amenability of a countable group, which illustrates importance and naturalness of this notion. We shall briefly mention just some of them, referring the reader to [Gre69], [Pie84] and [Pat88] for more details. Moreover, the notion of amenability has been extended to objects other than groups, in particular, to group actions, equivalence relations, and, more generally, to groupoids (see, e.g., [ADR00]).

The main application of the notion of amenability is its characterization as a *fixed point property.* Namely, *a countable group G is amenable if and only if any continuous affine action of G on a compact space has a fixed point.* An example of such an action arises in the following way. Let X be a compact topological space endowed with a continuous action of G, and let $\mathcal{P}(X)$ denote the space of probability measures on X endowed with the weak* topology. Then $\mathcal{P}(X)$ has a natural affine structure, and the action of G extends to a continuous affine action on $\mathcal{P}(X)$. Therefore, *any continuous action of an amenable group on a compact space has a finite invariant measure*[4] (in fact, this property can also be shown to be equivalent to amenability). In the case of the group of integers $\mathbb{Z}$ this result

moyennable in French), literally meaning "averageable". In English, however, Mahlon M. Day suggested to use (apparently, first as a pun) the word *amenable*, which appeared in print in this context for the first time in 1949 [Day49]. It is curious that Day himself, when he later described the history of this term in [Day83] on the occasion of the nomination of his paper [Day57] as a "Citation Classic" dated its appearance to 1955: *In* 1929*, von Neumann studied a new class of groups, those with invariant means on the bounded functions. My thesis* (1939) *studied semigroups with invariant means; thereafter, I worked in the field alternately with the geometry of Banach spaces. I finished a large geometrical project in* 1955 *and turned back to invariant means; in order to talk to my students I invented the term 'amenable* (*pronounced as amean'able*) *semigroups'.*

is known as the *Krylov–Bogolyubov theorem* [KB37] (which is one of the starting points of the modern theory of topological dynamical systems).

Yet another characterization of amenable groups can be given in terms of their *isoperimetric properties.* This condition is basically a specialization of Reiter's condition to sequences of measures of special kind (although historically it was introduced by Følner [Føl55] some 10 years before Reiter). Let A_n be a sequence of finite subsets of G, and let λ_n be the associated uniform probability measures on A_n. Then Reiter's condition for the sequence λ_n is equivalent to the following condition on the sets A_n:

$$\frac{|gA_n \triangle A_n|}{|A_n|} \underset{n\to\infty}{\longrightarrow} 0 \qquad \forall g \in G ,$$

where $\triangle$ denotes the symmetric difference of two sets, and $|A|$ is the cardinality of a finite set A. A sequence of sets A_n satisfying the above condition is called a *Følner sequence*, and the condition itself is called *Følner's condition.* Obviously, Følner's condition implies Reiter's condition; however, the usual "slicing" isoperimetric techniques also allow one to prove the converse, so that *Følner's condition is equivalent to amenability.*

For finitely generated groups Følner's condition takes especially simple form. Indeed, in this case it is enough to verify it for the elements g from a finite generating set $K \subset G$ only. Let us assume that K is symmetric, and denote by $\Gamma = \Gamma(G, K)$ the (left) *Cayley graph* of the group G determined by K (i.e., the vertex set is G, and the edges are of the form (g, kg) with $g \in G$ and $k \in K$)[5]. For a set $A \subset G$ denote by $\partial A = \partial_K A \subset A$ its *boundary* in the Cayley graph, i.e., the set of all points from A which have a neighbor from the complement of A. Then a sequence of sets $A_n \subset G$ is Følner if and only if

$$\frac{|\partial A_n|}{|A_n|} \underset{n\to\infty}{\longrightarrow} 0 .$$

Existence of a sequence of sets as above is an *isoperimetric characterization of amenability.*

Example 3. For the group $\mathbb{Z}$ with the standard generating set $\{\pm 1\}$ the boundary of the segment $A_n = \{0, 1, 2, \dots, n\}$ consists of two points $\{0, n\}$, whereas $|A_n| \to \infty$, so that $\{A_n\}$ is a Følner sequence.

2.C. Elementary groups

The class of amenable groups is closed with respect to the "elementary" operations of taking subgroups, quotients, extensions and inductive limits. Finite and abelian groups are amenable (cf. Example 2). The minimal class of groups containing finite

[4] The first proof of this fact by Bogolyubov [Bog39] published in 1939 (immediately after [KB37]) in a rather obscure journal in Ukrainian remained almost unknown, see [Ano94, CSGdlH99].

[5] Elsewhere in this paper we shall always deal with the *right* Cayley graphs. However, in order to keep the notations consistent, here it is more convenient to consider the left Cayley graphs.

and abelian groups and closed with respect to the above elementary operations is called *elementary amenable* (EA) [Day57].

In the above paper Day asked the question whether every amenable group is elementary amenable. The first example of an amenable but not elementary amenable group is the group of intermediate growth (see below) found by Grigorchuk [Gri80, Gri85]. Later, a finitely presented amenable extension of the Grigorchuk group was constructed in [Gri98].

However, there is yet another way to obtain "obviously amenable" groups. It is related with the notion of *growth*. Let G be a finitely generated group with a symmetric generating set K. Denote by B_n the n-ball of the Cayley graph metric on G centered at the group identity, or, in other words, the set of all elements of G which can be presented as products of not more than n generators from K. The sequence $|B_n|$ is submultiplicative, so that there exists a limit $\lim \log |B_n|/n$. The group G is said to have *exponential* or *subexponential growth* depending on whether this limit is positive or zero (this property does not depend on the choice of a generating set K).

The class of groups of subexponential growth contains all the groups of *polynomial growth* (the ones for which $|B_n|$ is bounded from above by a polynomial function; by a theorem of Gromov [Gro81] these are precisely finite extensions of nilpotent groups), but there are also examples of groups of *intermediate growth*, i.e., the ones whose growth is subexponential without being polynomial. First examples of this kind were constructed by Grigorchuk [Gri85], and these groups can often be realized as self-similar groups (see [BGN03]). For instance, the most famous of the Grigorchuk groups has 4 generators acting on the rooted binary tree with the matrix presentation

$$a \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} , \quad b \mapsto \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix} , \quad c \mapsto \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} , \quad d \mapsto \begin{pmatrix} 1 & 0 \\ 0 & b \end{pmatrix} .$$

If the growth of G is subexponential, then the sequence B_n necessarily contains a Følner subsequence, so that *the groups of subexponential growth are amenable.* Therefore, one can change the definition of elementary amenable groups by extending the set of "building blocks": the minimal class of groups containing finite, abelian and subexponential groups and closed with respect to the elementary operations is called *subexponentially elementary amenable* (SEA). Thus, a natural goal is to find amenable groups which are not subexponentially elementary (see [Gri98, CSGdlH99]).

It is the Basilica group $\mathcal{B}$ which provided the first example of this kind. It was shown in [GŻ02a] that it does not belong to the class SEA, whereas it was proved in [BV05] that the Basilica group is amenable. We shall now explain the role of random walks in the proof of amenability of $\mathcal{B}$.

Remark. It is worth mentioning at this point that *non-amenable groups do exist*, and there is actually quite a lot of them (for instance, numerous matrix groups: non-elementary Fuchsian and Kleinian groups, lattices in semi-simple Lie groups, etc.). The first example is of course the *free group* $\mathcal{F}_d$ with $d \geq 2$ generators (since any discrete group can be obtained from free groups by the elementary operations above, amenability of free groups would have implied amenability of *all* groups). Non-amenability of $\mathcal{F}_d$ may be explained in many different ways by using various definitions. For instance, it is not hard to build a *paradoxical decomposition* of $\mathcal{F}_d$ (see [CSGdlH99] and the references therein), which prevents it from having an invariant mean. Another way consists in noticing that there are continuous actions of $\mathcal{F}_d$ on compact sets admitting no finite invariant measures. Let, for instance $d = 2$. Then an action of $\mathcal{F}_2$ on a compact K is determined by specifying two homeomorphisms of K corresponding to the generators of $\mathcal{F}_2$. If K is the circle, then the only measure preserved by any irrational rotation is the Lebesgue measure. Take such a rotation for the first homeomorphism from the definition of the action. Then, if we choose the second homeomorphism in such a way that it does not preserve the Lebesgue measure, then these two homeomorphisms (therefore, the associated action of $\mathcal{F}_2$) do not have any invariant measure, so that $\mathcal{F}_2$ is not amenable.

3. Random walks

3.A. Convolution powers

The main idea behind the use of random walks for establishing amenability at first glance looks counterproductive. Let us replace arbitrary approximatively invariant sequences of probability measures λ_n on G from Reiter's condition with sequences of very special form, namely, with sequences of convolution powers μ^{*n} of a single probability measure μ on G. In the same way as with Følner's characterization of amenability (see Section 2.B), it turns out that this restricted class of approximatively invariant sequences is still sufficient in order to characterize amenability. More precisely, *a group G is amenable if and only if there exists a probability measure μ on G such that the sequence of its convolution powers μ^{*n} is approximatively invariant.* In one direction (the one we need for proving amenability) this is just a particular case of Reiter's condition, whereas in the opposite direction it was conjectured by Furstenberg [Fur73] and later independently proved by Kaimanovich–Vershik [VK79, KV83] and by Rosenblatt [Ros81].

Now, working with sequences of convolution powers instead of arbitrary sequences of probability measures on G is actually easier, because of their description as one-dimensional distributions of *random walks* on G. Moreover, one can use a very powerful quantitative criterion of whether a given sequence of convolution powers is approximatively invariant. It is provided by the *entropy theory of random walks* which we shall briefly describe below.

Random walks are in a sense the most homogeneous Markov chains: they are homogeneous both in space and in time. The latter property means that the assignment $x \mapsto \pi_x$ of transition probabilities to the points from the state space is equivariant with respect to a group action, and the simplest instance of such an action is, of course, the action of a group on itself.

More formally, the (right) *random walk* on a countable group G determined by a probability measure μ is the Markov chain with the state space G and the transition probabilities

$$\pi_g(g') = \mu(g^{-1}g')$$

equivariant with respect to the left action of the group on itself. In other words, from a point g the random walk moves at the next moment of time to the point gh, where the random *increment* h is chosen according to the distribution μ. We shall use for this description of transition probabilities of the random walk (G, μ) the notation

$$g \xmapsto[h\sim\mu]{} gh \ .$$

Thus, if the random walk starts at moment 0 from a point g_0, then its position at time n is

$$g_n = g_0 h_1 h_2 \dots h_n \ ,$$

where h_i is a *Bernoulli sequence* of independent μ-distributed increments. Therefore, the distribution of the position g_n of this random walk at time n is the translate $g_0\mu^{*n}$ of the *n-fold convolution power* μ^{*n} of the measure μ (the *convolution* of two probability measures μ_1, μ_2 on G is defined as the image of the product measure $\mu_1 \otimes \mu_2$ on $G \times G$ under the map $(g_1, g_2) \mapsto g_1 g_2$).

3.B. Trivial future

Reiter's condition

$$\|g\mu^{*n} - \mu^{*n}\| \xrightarrow[n\to\infty]{} 0 \qquad \forall\, g \in G$$

for the sequence of convolution powers μ^{*n} means that the time n one-dimensional distributions of the random walk issued from the group identity e and from an arbitrary point $g \in G$ asymptotically coincide. Probabilistically, the fact that the one-dimensional distributions of a Markov chain asymptotically do not depend on their starting points means that the "remote future" (behaviour at infinity) of the chain does not depend on its present, which, in view of the classical Kolmogorov argument means that the "remote future" must be trivial.

In order to explain the latter notion in a more rigorous way, let us look at two examples: the simple random walks on the 2-dimensional integer lattice ($\equiv$ the Cayley graph of the group $\mathbb{Z}^2$) and on the homogeneous tree of degree 4 ($\equiv$ the Cayley graph of the free group $\mathcal{F}_2$ with 2 generators). The simple random walk on a graph is the one whose transition probabilities are equidistributed among neighbours; the simple random walk on the Cayley graph of a group is precisely the random walk on this group determined by the probability measure $\mu = \mu_K$ equidistributed on the generating set K. Locally, each of these graphs is regular of degree 4 (each point has 4 neighbours), but their global geometry is very different, see Figure 5. In particular, the Cayley graph of $\mathcal{F}_2$ is endowed with a natural boundary $\partial\mathcal{F}_2$ which can, for instance, be identified with the space of all paths without backtracking issued from the identity of the group (or any other reference point).

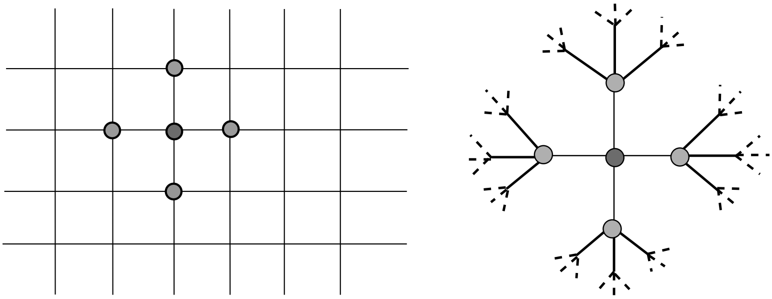

Figure 5

The global behaviour of sample paths of simple random walks on these graphs is also very different. Sample paths of the simple random walk on $\mathcal{F}_2$ converge a.s. to the boundary $\partial\mathcal{F}_2$ (of course, different sample paths may converges to different limits). Thus, these limit points can be used to distinguish sample paths by their behaviour at infinity. On the other hand, although sample paths of the random walk on $\mathbb{Z}^2$ are quite complicated (their scaling limit is the Brownian motion on the plane), they all look "the same", see Figure 6.

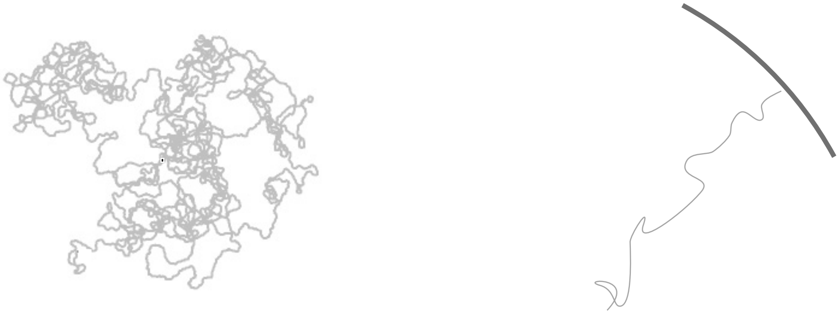

Figure 6

Formally speaking, the "remote future" of a Markov chain is described by its *tail σ-algebra*

$$\mathfrak{A}^\infty = \bigcap_n \mathfrak{A}_n^\infty ,$$

which is the limit of the decreasing sequence of σ-algebras $\mathfrak{A}_n^\infty$ generated by the positions of the chain at times $\geq n$. Thus, the boundary convergence of sample paths of the simple random walk on $\mathcal{F}_2$ at once implies non-triviality of its tail σ-algebra. On the other hand, in spite of absence of any visible behavior at infinity, proving triviality of the tail σ-algebra for the simple random walk on $\mathbb{Z}^2$ requires additional work.

On a formal level a criterion of the triviality of the tail σ-algebra of an arbitrary Markov chain is provided by the corresponding *0–2 law* [Der76, Kai92]. Its "zero part" for the random walk on a group G determined by a probability measure μ takes the following form (see [KV83]): the tail σ-algebra of the random walk is trivial if and only if

$$\|g\mu^{*n} - \mu^{*(n+1)}\| \underset{n\to\infty}{\longrightarrow} 0 \qquad \forall g \in \mathsf{supp}\,\mu \ .$$

Thus, *if the support of the measure μ generates G as a group, and the tail σ-algebra of the random walk (G,μ) is trivial, then* the sequence of Cesaro averages of the convolution powers satisfies Reiter's condition, and therefore *the group G is amenable* (actually, if μ is *aperiodic*, then the sequence of convolution powers μ^{*n} also satisfies Reiter's condition).

3.C. Asymptotic entropy

Let $p = (p_i)$ be a discrete probability distribution. By Shannon the *amount of information* contained in any outcome i is $-\log p_i$. The average amount of information per outcome

$$H(p) = -\sum p_i \log p_i$$

is called the *entropy* of the distribution p. The entropy has the following two properties (by which it is essentially characterized, see [YY83]):

- *monotonocity*: if p' is a quotient of the distribution p (i.e., p' is obtained by gluing together some of the states of p), then
$$H(p') < H(p) \ ,$$
- *additivity*: for any two distributions p_1, p_2 the entropy of their product is
$$H(p_1 \otimes p_2) = H(p_1) + H(p_2) \ .$$

Let μ be a probability measure on a countable group G with finite entropy $H(\mu)$. The (asymptotic) *entropy of the random walk* (G,μ) is defined as

$$h(G,\mu) = \lim_{n\to\infty} \frac{H(\mu^{*n})}{n} \ ,$$

in other words, $h(G,\mu)$ is the asymptotic mean specific amount of information about a single increment h_i in the product $g_n = h_1 h_2 \dots h_n$. Existence of the above limit follows from the fact that the sequence $H(\mu^{*n})$ is subadditive. Indeed, by the definition of convolution, the measure $\mu^{*(n+m)}$ is a quotient of the product measure $\mu^{*n} \otimes \mu^{*m}$, whence

$$H(\mu^{*(n+m)}) \le H(\mu^{*n} \otimes \mu^{*m}) = H(\mu^{*n}) + H(\mu^{*m})$$

by the above monotonicity and additivity properties.

The asymptotic entropy was first introduced by Avez [Ave72]. As it turned out, it plays a crucial role in understanding the asymptotic properties of the random walk (G,μ). Namely, as it was independently proved by Kaimanovich–Vershik

[VK79, KV83] and Derriennic [Der80], *the asymptotic entropy* $h(G,\mu)$ *vanishes if and only if the tail* σ*-algebra of the random walk* (G,μ) *is trivial.*

This result means that the "remote future" of the random walk (G,μ) is trivial if and only if the amount of information about the first increment h_1 in the random product $g_n = h_1 h_2 \dots h_n$ asymptotically vanishes. Indeed, on a more rigorous level, this amount of information is the *mean conditional entropy* (see [Roh67]) $H\left(h_1|g_n\right)$ of h_1 with respect to g_n. It can be easily seen to coincide with the difference $H\left(\mu^{*n}\right) - H\left(\mu^{*(n-1)}\right)$ between the entropies of two consecutive convolution powers, whence the claim.

3.D. Self-similarity and RWIDF

A *random walk with internal degrees of freedom* (RWIDF) on a group G is a Markov chain whose state space is the product of G by a certain space X (the *space of degrees of freedom*), and its transition probabilities are equivariant with respect to the action of G on itself. Thus, the transition probabilities are

$$(g,x) \xmapsto{\mu_{xy}(h)} (gh,y) \ ,$$

where $M = \{\mu_{xy} : x,y \in X\}$ is a $|X| \times |X|$ matrix of subprobability measures on G such that

$$\sum_y \|\mu_{xy}\| = 1 \qquad \forall x \in X \ .$$

The image of the RWIDF (G,M) under the map $(g,x) \mapsto x$ is the quotient Markov chain on X with the transition matrix

$$P = (p_{xy}) \ , \qquad p_{xy} = \|\mu_{xy}\| \ ,$$

which is the image of the matrix M under the augmentation map.

There is a natural link between random walks on self-similar groups and random walks with internal degrees of freedom. Let μ be a probability measure on a self-similar group G. Then the associated random walk is

$$g \xmapsto{\mu(h)} gh \ .$$

By using the self-similar embedding $G \hookrightarrow \mathsf{Sym}(X;G)$ it gives rise to the random walk on the generalized permutation group $\mathsf{Sym}(X;G)$ with the transition probabilities

$$M^g \xmapsto{\mu(h)} M^g M^h \ .$$

Since multiplication of the matrix $M^g \in \mathsf{Sym}(X;G)$ by the increment M^h is done row by row, we obtain the following Markov chain on the space of these rows:

$$R \xmapsto{\mu(h)} RM^h \ .$$

Due to the definition of the group $\mathsf{Sym}(X;G)$ the rows of the corresponding matrices can be identified with points of the product space $G \times X$ (each row has precisely one non-zero entry, so that it is completely described by the value of this entry and by its position). Therefore, the latter Markov chain can be interpreted as a Markov chain on $G \times X$ whose transition probabilities are easily seen to be invariant with respect to the left action of G on $G \times X$, i.e., as a random walk on G with internal degrees of freedom parameterized by the alphabet X. This RWIDF is described by the transition probabilities matrix

$$M^\mu = (\mu_{xy}) = \sum_g \mu(g) M^g \ .$$

Recall that stopping a Markov chain at the times when it visits a certain recurrent subset of the state space produces a new Markov chain on this recurrent subset (the *trace* of the original Markov chain). For a random walk with internal degrees of freedom on $G \times X$ determined by a matrix M take for such a subset the copy $G \times \{x\}$ of the group G obtained by fixing a value $x \in X$. It is recurrent provided the quotient chain on X is irreducible. The transition probabilities of the induced chain on $G \times \{x\}$ are obviously equivariant with respect to the left action of the group G on itself (because the original RWIDF also has this property). Therefore, the induced chain on $G \times \{x\}$ is actually the usual random walk determined by a certain probability measure μ^x on G.

The measures μ^x, $x \in X$ can be expressed in terms of the matrix M as

$$\begin{aligned}\mu^x &= \mu_{xx} + M_{x\overline{x}} \left(I + M_{\overline{xx}} + M_{\overline{xx}}^2 + \cdots\right) M_{\overline{x}x} \\ &= \mu_{xx} + M_{x\overline{x}} \left(I - M_{\overline{xx}}\right)^{-1} M_{\overline{x}x} \ ,\end{aligned}$$

where $M_{x\overline{x}}$ (resp., $M_{\overline{x}x}$) denotes the row $(\mu_{xy})_{y\neq x}$ (resp., the column $(\mu_{yx})_{y\neq x}$) of the matrix M with the removed element μ_{xx}, and $M_{\overline{xx}}$ is the $(d-1)\times(d-1)$ matrix (where $d = |X|$) obtained from M by removing its xth row and column. The multiplication above is understood in the usual matrix sense.[6]

This is elementary probability. We look at the quotient chain on X and replace its transition probabilities p_{xy} with the transition measures μ_{xy} in the identity

$$\begin{aligned}1 &= p_{xx} + \sum_{n=0}^{\infty} \sum_{y_0,\dots,y_n\neq x} p_{xy_0} p_{y_0y_1} \cdots p_{y_{n-1}y_n} p_{y_nx} \\ &= p_{xx} + P_{x\overline{x}} \left(I + P_{\overline{xx}} + P_{\overline{xx}}^2 + \cdots\right) P_{\overline{x}x} \\ &= p_{xx} + P_{x\overline{x}} \left(I - P_{\overline{xx}}\right)^{-1} P_{\overline{x}x} \ ,\end{aligned}$$

which yields

$$\begin{aligned}\mu^x &= \mu_{xx} + \sum_{n=0}^{\infty} \sum_{y_0,\dots,y_n\neq x} \mu_{xy_0} \mu_{y_0y_1} \cdots \mu_{y_{n-1}y_n} \mu_{y_nx} \\ &= \mu_{xx} + M_{x\overline{x}} \left(I + M_{\overline{xx}} + M_{\overline{xx}}^2 + \cdots\right) M_{\overline{x}x} \\ &= \mu_{xx} + M_{x\overline{x}} \left(I - M_{\overline{xx}}\right)^{-1} M_{\overline{x}x} \ .\end{aligned}$$

[6] As it is pointed out in [GN07], this formula corresponds to the classical operation of taking the *Schur complement* of a matrix.

The first term in this formula corresponds to staying at the point x (and performing on G the jump determined by the measure μ_{xx}), whereas in the second term the first factor corresponds to moving from x to $X \setminus \{x\}$, the second one to staying in $X \setminus \{x\}$ (each matrix power $M^n_{\overline{xx}}$ corresponding to staying in $X \setminus \{x\}$ for precisely n steps), and the third one to moving back from $X \setminus \{x\}$ to the point x. The matrix notation automatically takes care of what is going on with the G-component of the RWIDF.

The measures μ^x also admit the following interpretation in terms of the original random walk (G, μ) with the sample paths (g_n): we look at it only at the moments n when $g_n(T_x) = T_x$, and μ^x is then the law of the induced random walk on the group $\mathsf{Aut}(T_x) \cong \mathsf{Aut}(T)$.

3.E. The Münchhausen trick

Let us now look at what happens with the asymptotic entropy in the course of the transformations

$$\mu \mapsto M = M^\mu \mapsto \mu^x \ .$$

Since the information contained in a matrix does not exceed the sum of information about each row, passing to asymptotic entropies we obtain the inequality

$$h(G, \mu) \le dh(G, M) \ .$$

Further, all points $x \in X$ are visited by RWIDF (G, M) with the same asymptotic frequency $1/d$ (because the uniform distribution is stationary for the quotient chain on X), whence

$$h(G, \mu^x) = dh(G, M) \ge h(G, \mu) \ .$$

The idea of the *Münchhausen trick*[7] [Kai05] consists in combining two observations. The first one is the above entropy inequality. The second observation is that if the measure μ^x is a non-trivial convex combination

$$\mu^x = (1 - \alpha)\delta_e + \alpha\mu \ , \qquad 0 < \alpha < 1$$

of the original measure μ and the δ-measure at the identity of the group (in which case we call the measure μ *self-similar*), then

$$h(G, \mu^x) = \alpha h(G, \mu) \ .$$

The result of these two observations is the inequality

$$h(G, \mu) \le \alpha h(G, \mu) \ .$$

Taken into account that $0 < \alpha < 1$, it is only possible if $h(G, \mu) = 0$, which proves amenability of the group G.

We shall now show how the Münchhausen trick works in two particular cases: for the Basilica group $\mathcal{B}$ and for the Mother groups $\mathfrak{M}$.

[7] It is named after venerable Baron Münchhausen who *"... once rode on a cannon-ball and next told about it. Another time he reported he had to get himself and the good horse he sat on, out of a quagmire by pulling his own hair till he saved himself and his horse... "*

3.F. The Basilica group

As we have seen in Section 1.D, the Basilica group is defined by the matrix presentation

$$\mathcal{B} : a \mapsto \begin{pmatrix} b & 0 \\ 0 & 1 \end{pmatrix} , \qquad b \mapsto \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix} .$$

Take on $\mathcal{B}$ a symmetric probability measure μ supported by the generators a, b and their inverses[8]:

$$\mu = \alpha \left(a + a^{-1} \right) + \beta \left(b + b^{-1} \right) ,$$

for $\alpha, \beta > 0$ such that $2(\alpha + \beta) = 1$. The matrix associated with the measure μ is

$$\begin{aligned} M^\mu &= \alpha M^a + \alpha M^{a^{-1}} + \beta M^b + \beta M^{b^{-1}} \\ &= \alpha \begin{pmatrix} b & 0 \\ 0 & 1 \end{pmatrix} + \alpha \begin{pmatrix} b^{-1} & 0 \\ 0 & 1 \end{pmatrix} + \beta \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ a^{-1} & 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha(b + b^{-1}) & \beta(1 + a) \\ \beta(1 + a^{-1}) & 2\alpha \end{pmatrix} . \end{aligned}$$

Therefore, the trace of the RWIDF $(\mathcal{B}, M^\mu)$ on the copy of $\mathcal{B}$ corresponding to the first letter 1 of the 2-letter alphabet $\{1, 2\}$ is the random walk governed by the measure

$$\begin{aligned} \widetilde{\mu} = \widetilde{\mu}^1 &= \mu_{11} + \mu_{12} \left(1 - \mu_{22}\right)^{-1} \mu_{21} \\ &= \alpha(b + b^{-1}) + \frac{\beta}{2}(1 + a)(1 + a^{-1}) = \beta + \frac{\beta}{2}(a + a^{-1}) + \alpha(b + b^{-1}) . \end{aligned}$$

If

$$\frac{\alpha}{\beta} = \frac{\beta}{2\alpha} \iff 2\alpha^2 = \beta^2 ,$$

then

$$\widetilde{\mu} = \beta + (1 - \beta)\mu ,$$

so that the measure μ is self-similar, and Münchhausen's trick is applicable.

3.G. The Mother group

For proving amenability of the Mother group $\mathfrak{M}$ (here and below we omit the alphabet X) one can apply an approach somewhat different from the one which was used above for the Basilica group. It is based on the fact that the Mother group $\mathfrak{M}$ is generated by two finite subgroups A and B. We take for the measure μ the convolution product of the uniform measures m_A and m_B on these subgroups:

$$\mu = m_A * m_B .$$

Then the matrix M^μ has a very special form

$$M^\mu = M^{\mu_A} M^{\mu_B} = E_d \begin{pmatrix} \mu_B & 0 \\ 0 & \mu_A E_{d-1} \end{pmatrix} ,$$

[8] For simplicity we pass from now on to the group algebra notations putting g for a δ-measure δ_g, $g \neq e$ and just 1 for the δ-measure δ_e concentrated at the identity of the group.

where $d = |X|$, and E_d denotes the order d matrix with entries $1/d$, so that M^μ has identical rows with entries

$$M^\mu_{xy} = \begin{cases} \mu_B/d & \text{if } y = o \,, \\ \mu_A/d & \text{otherwise} \,. \end{cases}$$

It means that transition probabilities of the associated RWIDF $(\mathfrak{M} \times X, M^\mu)$ do not depend on x, so that its projection to $\mathfrak{M}$ is just the random walk $(\mathfrak{M}, \widetilde{\mu})$ determined by the measure

$$\widetilde{\mu} = \sum_y M^\mu_{xy} = \frac{d-1}{d}\mu_A + \frac{1}{d}\mu_B \,,$$

whereas the projection of RWIDF $(\mathfrak{M}\times X, M^\mu)$ to X is the sequence of independent X-valued random variables with uniform distribution on X (because all entries M^μ_{xy} have mass $1/d$). Thus,

$$h(\mathfrak{M}, \mu) \le d\, h(\mathfrak{M}, \widetilde{\mu}) \,.$$

The measure $\widetilde{\mu}$ is a convex combination of the idempotent measures μ_A and μ_B, so that its convolution powers are essentially convex combinations of the convolution powers of μ. The total number of m_B's in the n-fold convolution of $\widetilde{\mu}$ is $\sim n/d$, but some of them disappear because $m_B m_B = m_B$, so that in fact $\widetilde{\mu}^{*n}$ is the convolution of about $\frac{d-1}{d^2}n$ copies of $\mu = m_A * m_B$. Thus,

$$h(\mathfrak{M}, \widetilde{\mu}) = \frac{d-1}{d^2} h(\mathfrak{M}, \mu) \,,$$

whence $h(\mathfrak{M}, \mu) = 0$.

References

[ADR00] Claire Anantharaman-Delaroche and Jean Renault, *Amenable groupoids*, Monographies de L'Enseignement Mathématique [Monographs of L'Enseignement Mathématique], vol. 36, L'Enseignement Mathématique, Geneva, 2000, With a foreword by Georges Skandalis and Appendix B by E. Germain. MR 2001m:22005.

[Ano94] D.V. Anosov, *On N.N. Bogolyubov's contribution to the theory of dynamical systems*, Uspekhi Mat. Nauk **49** (1994), no. 5(299), 5–20. MR 1311227 (96a:01029).

[Ave72] André Avez, *Entropie des groupes de type fini*, C. R. Acad. Sci. Paris Sér. A-B **275** (1972), A1363–A1366. MR 0324741 (48 #3090).

[BG00] L. Bartholdi and R.I. Grigorchuk, *On the spectrum of Hecke type operators related to some fractal groups*, Tr. Mat. Inst. Steklova **231** (2000), no. Din. Sist., Avtom. i Beskon. Gruppy, 5–45. MR 1841750 (2002d:37017).

[BG02] Laurent Bartholdi and Rostislav Grigorchuk, *On a group associated to z^2-1*, arXiv: math.GR/0203244, 2002.

[BGN03] Laurent Bartholdi, Rostislav Grigorchuk, and Volodymyr Nekrashevych, *From fractal groups to fractal sets*, Fractals in Graz 2001, Trends Math., Birkhäuser, Basel, 2003, pp. 25–118. MR 2091700 (2005h:20056).

[Bie90] Ben Bielefeld (ed.), *Conformal dynamics problem list*, Institute for Mathematical Sciences preprint series, no. 90-1, SUNY Stony Brook, 1990, arXiv: math.DS/9201271.

[BKN08] Laurent Bartholdi, Vadim Kaimanovich, and Volodymyr Nekrashevych, *On amenability of automata groups*, arXiv:0802.2837, 2008.

[BN03] E. Bondarenko and V. Nekrashevych, *Post-critically finite self-similar groups*, Algebra Discrete Math. (2003), no. 4, 21–32. MR 2070400 (2005d:20041).

[BN06] Laurent Bartholdi and Volodymyr Nekrashevych, *Thurston equivalence of topological polynomials*, Acta Math. **197** (2006), no. 1, 1–51. MR 2285317 (2008c:37072).

[Bog39] N.N. Bogolyubov, *On some ergodic properties of continuous transformation groups*, Nauch. Zap. Kiev Univ. Phys.-Mat. Sb. **4** (1939), no. 5, 45–52, in Ukranian, also *Selected works in mathematics*, Fizmatlit, Moscow, 2006, pp. 213–222 (in Russian).

[BS98] A.M. Brunner and Said Sidki, *The generation of* $\mathrm{GL}(n,\mathbb{Z})$ *by finite state automata*, Internat. J. Algebra Comput. **8** (1998), no. 1, 127–139. MR 1492064 (99f:20055).

[BV05] Laurent Bartholdi and Bálint Virág, *Amenability via random walks*, Duke Math. J. **130** (2005), no. 1, 39–56. MR 2176547 (2006h:43001).

[CSGdlH99] Tullio Ceccherini-Silberstein, Rostislav I. Grigorchuk, and Pierre de la Harpe, *Amenability and paradoxical decompositions for pseudogroups and discrete metric spaces*, Proc. Steklov Inst. Math. **224** (1999), no. 1, 57–97. MR 1721355 (2001h:43001).

[CW89] A. Connes and E.J. Woods, *Hyperfinite von Neumann algebras and Poisson boundaries of time dependent random walks*, Pacific J. Math. **137** (1989), no. 2, 225–243. MR 90h:46100.

[Day49] Mahlon M. Day, *Means on semigroups and groups*, Bull. Amer. Math. Soc. **55** (1949), 1054–1055, abstract 55–11-507.

[Day57] ———, *Amenable semigroups*, Illinois J. Math. **1** (1957), 509–544. MR 0092128 (19,1067c).

[Day83] ———, *Citation classic-amenable semigroups*, Current Contents Phys. Chem. Earth (1983), no. 26, 18–18.

[Der76] Yves Derriennic, *Lois "zéro ou deux" pour les processus de Markov. Applications aux marches aléatoires*, Ann. Inst. H. Poincaré Sect. B (N.S.) **12** (1976), no. 2, 111–129. MR 54 #11508.

[Der80] ———, *Quelques applications du théorème ergodique sous-additif*, Conference on Random Walks (Kleebach, 1979) (French), Astérisque, vol. 74, Soc. Math. France, Paris, 1980, pp. 183–201, 4. MR 588163 (82e:60013).

[dlH00] Pierre de la Harpe, *Topics in geometric group theory*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL, 2000. MR 1786869 (2001i:20081).

[Fal03] Kenneth Falconer, *Fractal geometry*, second ed., John Wiley & Sons Inc., Hoboken, NJ, 2003, Mathematical foundations and applications. MR 2118797 (2006b:28001).

[Føl55] Erling Følner, *On groups with full Banach mean value*, Math. Scand. **3** (1955), 243–254. MR 0079220 (18,51f).

[Fur73] Harry Furstenberg, *Boundary theory and stochastic processes on homogeneous spaces*, Harmonic analysis on homogeneous spaces (Proc. Sympos. Pure Math., Vol. XXVI, Williams Coll., Williamstown, Mass., 1972), Amer. Math. Soc., Providence, R.I., 1973, pp. 193–229. MR 50 #4815.

[GM05] Yair Glasner and Shahar Mozes, *Automata and square complexes*, Geom. Dedicata **111** (2005), 43–64. MR 2155175 (2006g:20112).

[GN07] Rostislav Grigorchuk and Volodymyr Nekrashevych, *Self-similar groups, operator algebras and Schur complement*, J. Mod. Dyn. **1** (2007), no. 3, 323–370. MR 2318495 (2008e:46072).

[Gre69] Frederick P. Greenleaf, *Invariant means on topological groups and their applications*, Van Nostrand Mathematical Studies, No. 16, Van Nostrand Reinhold Co., New York, 1969. MR 40 #4776.

[Gri80] R.I. Grigorchuk, *On Burnside's problem on periodic groups*, Funktsional. Anal. Appl. **14** (1980), no. 1, 41–43. MR 565099 (81m:20045).

[Gri85] ———, *Degrees of growth of finitely generated groups and the theory of invariant means*, Math. SSSR Izv. **25** (1985), no. 2, 259–300. MR 764305 (86h:20041).

[Gri98] ———, *An example of a finitely presented amenable group that does not belong to the class EG*, Sb. Math. **189** (1998), no. 1-2, 75–95. MR 1616436 (99b:20055).

[Gro81] Mikhael Gromov, *Groups of polynomial growth and expanding maps*, Inst. Hautes Études Sci. Publ. Math. (1981), no. 53, 53–73. MR 623534 (83b:53041).

[GŻ02a] Rostislav I. Grigorchuk and Andrzej Żuk, *On a torsion-free weakly branch group defined by a three state automaton*, Internat. J. Algebra Comput. **12** (2002), no. 1-2, 223–246, International Conference on Geometric and Combinatorial Methods in Group Theory and Semigroup Theory (Lincoln, NE, 2000). MR 2003c:20048.

[GŻ02b] ———, *Spectral properties of a torsion-free weakly branch group defined by a three state automaton*, Computational and statistical group theory (Las Vegas, NV/Hoboken, NJ, 2001), Contemp. Math., vol. 298, Amer. Math. Soc., Providence, RI, 2002, pp. 57–82. MR 2003h:60011.

[Kai92] Vadim A. Kaimanovich, *Measure-theoretic boundaries of Markov chains,* 0-2 *laws and entropy*, Harmonic analysis and discrete potential theory (Frascati, 1991), Plenum, New York, 1992, pp. 145–180. MR 94h:60099.

[Kai95] ———, *The Poisson boundary of covering Markov operators*, Israel J. Math. **89** (1995), no. 1-3, 77–134. MR 96k:60194.

[Kai03] ———, *Random walks on Sierpiński graphs: hyperbolicity and stochastic homogenization*, Fractals in Graz 2001, Trends Math., Birkhäuser, Basel, 2003, pp. 145–183. MR 2091703 (2005h:28022).

[Kai05] ———, *“Münchhausen trick” and amenability of self-similar groups*, Internat. J. Algebra Comput. **15** (2005), no. 5-6, 907–937. MR 2197814.

[KB37] Nicolas Kryloff and Nicolas Bogoliouboff, *La théorie générale de la mesure dans son application à l’étude des systèmes dynamiques de la mécanique non linéaire*, Ann. of Math. (2) **38** (1937), no. 1, 65–113. MR 1503326.

[Kig01] Jun Kigami, *Analysis on fractals*, Cambridge Tracts in Mathematics, vol. 143, Cambridge University Press, Cambridge, 2001. MR 1840042 (2002c:28015).

[KS83] András Krámli and Domokos Szász, *Random walks with internal degrees of freedom. I. Local limit theorems*, Z. Wahrsch. Verw. Gebiete **63** (1983), no. 1, 85–95. MR 85f:60098.

[KV83] V.A. Kaimanovich and A.M. Vershik, *Random walks on discrete groups: boundary and entropy*, Ann. Probab. **11** (1983), no. 3, 457–490. MR 85d:60024.

[Nek05] Volodymyr Nekrashevych, *Self-similar groups*, Mathematical Surveys and Monographs, vol. 117, American Mathematical Society, Providence, RI, 2005. MR 2162164 (2006e:20047).

[Nek08] ———, *Free subgroups in groups acting on rooted trees*, arXiv:0802.2554, 2008.

[NT08] Volodymir Nekrashevych and Alexander Teplyaev, *Groups and analysis on fractals*, Analysis on Graphs and its Applications (Proc. Sympos. Pure Math., Vol. 77), Amer. Math. Soc., Providence, R.I., 2008, pp. 143–180.

[Pat88] Alan L.T. Paterson, *Amenability*, Mathematical Surveys and Monographs, vol. 29, American Mathematical Society, Providence, RI, 1988. MR 90e:43001.

[Pie84] Jean-Paul Pier, *Amenable locally compact groups*, Pure and Applied Mathematics, John Wiley & Sons Inc., New York, 1984, A Wiley-Interscience Publication. MR 86a:43001.

[Rei65] H. Reiter, *On some properties of locally compact groups*, Nederl. Akad. Wetensch. Proc. Ser. A 68=Indag. Math. **27** (1965), 697–701. MR 0194908 (33 #3114).

[Roh67] V.A. Rohlin, *Lectures on the entropy theory of transformations with invariant measure*, Uspehi Mat. Nauk **22** (1967), no. 5 (137), 3–56. MR 0217258 (36 #349).

[Ros81] Joseph Rosenblatt, *Ergodic and mixing random walks on locally compact groups*, Math. Ann. **257** (1981), no. 1, 31–42. MR 83f:43002.

[RT09] Luke G. Rogers and Alexander Teplyaev, *Laplacians on the Basilica Julia set*, Commun. Pure Appl. Anal. (2009), to appear.

[Sid00] Said Sidki, *Automorphisms of one-rooted trees: growth, circuit structure, and acyclicity*, J. Math. Sci. (New York) **100** (2000), no. 1, 1925–1943, Algebra, 12. MR 1774362 (2002g:05100).

[VK79] A.M. Vershik and V.A Kaimanovich, *Random walks on groups: boundary, entropy, uniform distribution*, Dokl. Akad. Nauk SSSR **249** (1979), no. 1, 15–18. MR 553972 (81f:60098).

[vN29] John von Neumann, *Zur allgemeinen Theorie des Maßes*, Fund. Math. **13** (1929), 73–116 and 333, also *Collected works*, vol. I, pages 599–643.

[VV07] Mariya Vorobets and Yaroslav Vorobets, *On a free group of transformations defined by an automaton*, Geom. Dedicata **124** (2007), 237–249. MR 2318547 (2008i:20030).

[YY83] A.M. Yaglom and I.M. Yaglom, *Probability and information*, Theory and Decision Library, vol. 35, D. Reidel Publishing Co., Dordrecht, 1983, Translated from the third Russian edition by V. K. Jain. MR 736349 (85d:94001).

Vadim A. Kaimanovich
Jacobs University Bremen
D-28759 Bremen, Germany
e-mail: `vadim.kaimanovich@gmail.com`

Part 2

Conformal Dynamics and Schramm-Loewner Evolution

Progress in Probability, Vol. 61, 73–107

Multifractal Analysis of the Reverse Flow for the Schramm-Loewner Evolution

Gregory F. Lawler

Dedicated to the memory of Oded Schramm
without whom this paper would not exist.

Abstract. The Schramm-Loewner evolution (SLE) is a one-parameter family of conformally invariant processes that are candidates for scaling limits for two-dimensional lattice models in statistical physics. Analysis of SLE curves requires estimating moments of derivatives of random conformal maps. We show how to use the Girsanov theorem to study the moments for the reverse Loewner flow. As an application, we give a new proof of Beffara's theorem about the dimension of SLE curves.

Mathematics Subject Classification (2000). Primary 60J60;
Secondary 37E35, 82B27.

Keywords. Schramm-Loewner evolution, multifractal, Hausdorff dimension.

1. Introduction

The Schramm-Loewner evolution (SLE) was introduced by Oded Schramm [11] as a candidate for scaling limits of models in statistical physics. It has led to a much greater rigorous understanding of scaling limits of critical models in two-dimensional statistical physics.

Here we give a brief introduction to SLE. See [5] for more details. Chordal SLE_κ in the upper half-plane $\mathbb{H}$ is defined in terms of conformal maps g_t defined by

$$\partial_t g_t(z) = \frac{a}{g_t(z) - V_t}, \quad g_0(z) = z, \tag{1}$$

where $a = 2/\kappa$ and V_t is a one-dimensional Brownian motion. There is a corresponding random curve $\gamma(0,t]$. The relation between the two is that for a fixed

Research supported by National Science Foundation grant DMS-0734151.

time t, if H_t denotes the unbounded component of $\mathbb{H}\setminus\gamma(0,t]$, then g_t is a conformal transformation of H_t onto $\mathbb{H}$ satisfying

$$g_t(z) = z + \frac{at}{z} + O(|z|^{-2}), \quad z \to \infty.$$

We let $f_t = g_t^{-1}$, which is a conformal transformation of $\mathbb{H}$ onto H_t, and $\hat{f}_t(z) = f_t(z + V_t)$. (We have chosen a particular parametrization of the SLE path. In the original definition, the Loewner equation was written with 2 replacing a, and then the function V_t was a Brownian motion with variance $\kappa = 2/a$. Our choice is a simple linear reparametrizaton, i.e., using a different unit of time. We write many of our formulas in terms of a but throughout the paper $a = 2/\kappa$.)

A number of problems in SLE lead to studying moments of derivatives. When considering moments of $|\hat{f}_t'(z)|$, one can instead consider a reverse-time Loewner flow. In this paper, we consider solutions of the time-reversed Loewner equation

$$\partial_t h_t(z) = \frac{a}{U_t - h_t(z)}, \quad h_0(z) = z, \tag{2}$$

where $U_t = -B_t$ is a standard Brownian motion. Using only the Loewner equation (1) and the time-reversibility of Brownian motion (see Section 10.3), one can show that the distribution of $h_t(z) - U_t$ is the same as the distribution of $\hat{f}_t(z)$. In many ways the reverse Loewner flow (2) is easier to analyze than the forward flow (1), see, e.g., [10, 4, 9].

In this paper we will study the moments of $|h_t'|$ and show how they can be studied using relatively standard methods of stochastic calculus. This builds on previous work, especially that of Rohde and Schramm [10] who first studied the moments in order to show that the curve γ exists. Moments have also been studied by a number of other authors, see, e.g., [2]. There are several reasons to include a self-contained treatment of these moments. First, the recent work of the author [7, 8] relies on these estimates. Second, this is a nice example of a multifractal analysis that can be done. It illustrates the general technique in SLE of trying to reduce problems of the flow to a one-variable SDE and then to analyze the SDE. A particular emphasis is the role of the Girsanov theorem in analyzing the SDE.

Our approach to studying the moments is to find an appropriate martingale and then to use the Girsanov theorem to understand the distribution when weighted by the martingale. We first note that the scaling properties of Brownian motion imply that for each $r > 0$, the distribution of the random function $z \mapsto r^{-1}\, h_{r^2 t}(rz)$ is the same as the distribution of $z \mapsto h_t(z)$. In particular, the distribution of $h'_{r^2 t}(rz)$ is the same as that of $h_t'(z)$.

As an application, in Section 10 we give a new proof of Beffara's theorem [1] that the Hausdorff dimension of SLE_κ curves is $1 + \frac{\kappa}{8}$ for $\kappa \le 8$.

I would like to thank Tom Alberts for his detailed comments on an earlier version of this paper.

2. Studying the flow

If $z = x_z + iy_z \in \mathbb{H}$, let

$$Z_t(z) = X_t(z) + iY_t(z) = h_t(z) - U_t.$$

Then (2) can be written as

$$dZ_t(z) = -\frac{a}{Z_t(z)}\,dt + dB_t, \qquad Z_0(z) = z,$$

or

$$dX_t(z) = -\frac{a\,X_t(z)}{|Z_t(z)|^2}\,dt + dB_t, \quad \partial_t Y_t(z) = \frac{a\,Y_t(z)}{|Z_t(z)|^2}. \tag{3}$$

We write d for stochastic differentials (with respect to time) and ∂_t for actual derivatives. We note the following properties.

- $Y_t(z)$ increases with t and hence the solution to (2) exists for all times. Moreover, $\partial_t[Y_t(z)^2] \le 2a$ and hence
$$y_z^2 \le Y_t(z)^2 \le y_z^2 + 2at.$$
- For each $t \ge 0$, h_t is a conformal transformation of $\mathbb{H}$ onto a subdomain $h_t(\mathbb{H})$ satisfying
$$h_t(z) = z - \frac{at}{z} + O(|z|^{-2}), \quad z \to \infty.$$

We let

$$S_t(z) = \sin\left[\arg Z_t(z)\right] = \left[\frac{X_t(z)^2}{Y_t(z)^2} + 1\right]^{-1/2}, \quad \Psi_t(z) = \frac{|h_t'(z)|}{Y_t(z)}.$$

A calculation using Itô's formula gives

$$d\left[S_t(z)^{r/2}\right] = S_t(z)^{r/2}\left[\frac{(2ar + \frac{r^2}{2} + \frac{r}{2})X_t(z)^2 - \frac{r}{2}\,Y_t(z)^2}{|Z_t(z)|^4}\,dt - \frac{rX_t(z)}{|Z_t(z)|^2}\,dB_t\right]. \tag{4}$$

By differentiating (2) with respect to z, we can see that

$$\partial_t[\log h_t'(z)] = \frac{a}{Z_t(z)^2}.$$

Therefore,

$$\partial_t|h_t'(z)| = |h_t'(z)|\,\mathrm{Re}\left[\frac{a}{Z_t(z)^2}\right] = |h_t'(z)|\,\frac{a\left[X_t(z)^2 - Y_t(z)^2\right]}{|Z_t(z)|^4}, \tag{5}$$

and

$$\partial_t\Psi_t(z) = \Psi_t(z)\,\frac{-2a\,Y_t(z)^2}{|Z_t(z)|^4}.$$

In particular, $\Psi_t(z)$ decreases with t which implies

$$|h_t'(z)| \le \frac{Y_t(z)}{Y_0(z)} \le \sqrt{1 + 2a(t/y_z^2)}. \tag{6}$$

The next proposition introduces the family of martingales indexed by $r \in \mathbb{R}$ that will be the main tool for estimating the moments of $|h_t'(z)|$.

Proposition 2.1. *If $r \in \mathbb{R}$, $z = x_z + iy_z \in \mathbb{H}$, and*

$$\lambda = \lambda(r) = r\left(1 + \frac{1}{2a}\right) - \frac{r^2}{4a}, \qquad \zeta = \zeta(r) = r - \frac{r^2}{4a} = \lambda - \frac{r}{2a} \tag{7}$$

then

$$M_t(z) = |h_t'(z)|^{\lambda}\, Y_t(z)^{\zeta} S_t(z)^{-r} \tag{8}$$

is a martingale satisfying

$$dM_t(z) = \frac{r\, X_t(z)}{|Z_t(z)|^2}\, M_t(z)\, dB_t. \tag{9}$$

In particular,

$$\mathbb{E}\,[M_t(z)] = M_0(z) = y_z^{\zeta}\,[(x_z/y_z)^2 + 1]^{r/2},$$

Proof. The product rule combined with (3), (4), and (5) shows that $M_t = M_t(z)$ is a nonnegative local martingale satisfying (9). (Note that $|h_t'(z)|^{\lambda}, Y_t(z)^{\zeta}$ are differentiable quantities so there are no covariation terms.) We can use the Girsanov theorem to conclude that M_t is a martingale for $z \notin \mathbb{R}$. We give the sketch of the argument here. Readers unfamiliar with the use of stopping times with the Girsanov theorem for continuous local martingales should consult the appendix for more details.

If we use Girsanov's theorem and weight B_t by the local martingale $M_t(z)$ then

$$dB_t = \frac{r\, X_t(z)}{|Z_t(z)|^2}\, dt + d\tilde{B}_t$$

where $\tilde{B}_t$ is a Brownian motion with respect to the new measure $\mathbf{Q}$. In other words,

$$dX_t(z) = \frac{(r-a)\, X_t(z)}{|Z_t(z)|^2}\, dt + d\tilde{B}_t. \tag{10}$$

Note that $Y_t(z)$ and $|h_t'(z)|^d$ are differentiable quantities so their equations do not change in the new measure. (Actually the equation of $X_t(z)$ also does not change. What changes is the distribution of the random process B_t. In order to write the equation for $X_t(z)$ in terms of a Brownian motion in the new measure, the drift term is changed.) By comparing to a Bessel equation, it is easy to check that if $X_t(z)$ satisfies (10) then there is no explosion in finite time. Using (9), we see that $M_t(z)$ also does not have explosion in finite time. □

3. Multifractal analysis

Multifractal analysis refers to the study of moments of a random variable and measures obtained by weighting by powers of random variables. There is a significant overlap between this and large deviation theory. Here we discuss a simple version of this that applies to our situation.

Suppose we have a collection of random variables $D_t, t > 0$. The start of large deviation analysis is to estimate the exponential moments, e.g., to find a function $\zeta(\lambda)$ such that

$$\mathbb{E}[e^{\lambda D_t}] \approx e^{-\zeta(\lambda) t},$$

where $\approx$ means

$$\zeta(\lambda) = - \lim_{t \to \infty} \frac{\log \mathbb{E}[e^{\lambda D_t}]}{t}. \tag{11}$$

Many papers in probability are devoted to finding the function ζ for some particular random variables. While one often can only prove a result such as (11), there are many cases where one can give a stronger estimate:

$$\mathbb{E}[e^{\lambda D_t}] \asymp e^{-\zeta(\lambda) D_t}, \tag{12}$$

where the implicit constants in the $\asymp$ notation can be chosen uniformly over λ in an interval. If we know (12) and can show that ζ is C^2, then, as the next proposition shows, it is easy to conclude that (roughly speaking) the expectation in (12) concentrates on an event on which

$$D_t = -\zeta'(\lambda)\, t + O(t^{1/2}).$$

Proposition 3.1. *Suppose $\lambda_0 \in \mathbb{R}, \epsilon > 0$, $0 < c_1 < c_2 < \infty$, and ζ is a differentiable function such that for all $t \geq 1/\epsilon^2$,*

$$c_1\, e^{-t\zeta(\lambda)} \leq \mathbb{E}[e^{\lambda D_t}] \leq c_2\, e^{-t\zeta(\lambda)}, \quad |\lambda - \lambda_0| \leq \epsilon.$$

Suppose there exists $\alpha < \infty$ such that

$$|\zeta(\lambda) - \zeta(\lambda_0) - \zeta'(\lambda_0)\,(\lambda - \lambda_0)| \leq \alpha\,(\lambda - \lambda_0)^2, \quad |\lambda - \lambda_0| \leq \epsilon.$$

Then for all $t \geq 1/\epsilon^2$ and all $k > 0$,

$$\mathbb{E}\left[e^{\lambda_0 D_t}; |D_t - \mu t| \geq k\, t^{1/2}\right] \leq c_*\, e^{-k}\, \mathbb{E}\left[e^{\lambda_0 D_t}\right].$$

where $\mu = -\zeta'(\lambda_0), c_* = 2e^{\alpha} c_2/c_1$.

Proof.

$$\begin{aligned}
&\mathbb{E}[e^{\lambda_0 D_t}; D_t \geq \mu t + k t^{1/2}] \\
&\quad \leq e^{-t^{-1/2}(\mu t + k t^{1/2})}\, \mathbb{E}[e^{(\lambda_0 + t^{-1/2}) D_t}; D_t \geq \mu t + k t^{1/2}] \\
&\quad \leq c_2\, e^{-k}\, e^{-\mu t^{1/2}}\, \exp\{-t\zeta(\lambda_0 + t^{-1/2})\} \\
&\quad \leq c_2\, e^{-k}\, e^{\alpha}\, e^{-t\zeta(\lambda_0)} \leq (c_*/2)\, e^{-k}\, \mathbb{E}[e^{\lambda_0 D_t}],
\end{aligned}$$

Similarly, one shows that

$$\mathbb{E}[e^{\lambda_0 D_t}; D_t \leq \mu t - k t^{1/2}] \leq (c_*/2)\, e^{-k}\, \mathbb{E}[e^{\lambda_0 D_t}]. \qquad \square$$

Let $\mathbb{P}_t$ denote the probability measure with

$$\frac{d\mathbb{P}_t}{d\mathbb{P}} = \frac{e^{\lambda_0 D_t}}{\mathbb{E}[e^{\lambda_0 D_t}]}.$$

Then the conclusion of the proposition can be written as

$$\mathbb{P}_t\left\{\frac{|D_t - \mu t|}{\sqrt{t}} \geq k\right\} \leq c_* \, e^{-k}, \quad t \geq \epsilon^{-2}.$$

4. Moments of $|h'|$

In this section we fix $z = i$ and write $M_t, X_t, Y_t, \ldots$ for $M_t(i), X_t(i), Y_t(i), \ldots$ We will study the multifractal behavior of $D_t = \log|h'_{e^{2t}}(i)|$, i.e., we will find the function $\lambda \mapsto \zeta^*(\lambda)$ such that

$$\mathbb{E}[|h'_{t^2}(i)|^\lambda] \approx t^{-\zeta^*(\lambda)}.$$

Let

$$r_c = 2a + \frac{1}{2}, \quad \lambda_c := \lambda(r_c) = a + \frac{3}{16a} + 1, \quad \zeta_c := \zeta(r_c) = a - \frac{1}{16a}.$$

We show the significance of these values below; it is the value of λ such that (15) holds. If $r \leq r_c$, we can solve the quadratic equation (7), to get

$$r = r(\lambda) = 2a + 1 - \sqrt{(2a+1)^2 - 4a\lambda}, \quad \lambda \leq \lambda_c. \tag{13}$$

We can write ζ as a function of λ,

$$\zeta(\lambda) = \lambda - \frac{r}{2a} = \lambda + \frac{1}{2a}\sqrt{(2a+1)^2 - 4a\lambda} - 1 - \frac{1}{2a}. \tag{14}$$

As r increases from $-\infty$ to r_c, λ increases from $-\infty$ to λ_c. For $\lambda \leq \lambda_c$, we can write the martingale from Proposition 2.1 as

$$M_t = |h'_t(z)|^\lambda \, Y_t^{\zeta(\lambda)} \, S_t^{-r(\lambda)}.$$

Note that

$$\zeta'(\lambda) := \partial_\lambda \zeta(\lambda) = 1 - \frac{1}{\sqrt{(2a+1)^2 - 4a\lambda}}.$$

In particular, ζ is strictly concave with $\zeta'(-\infty) = 1$. The critical value λ_c satisfies

$$\zeta'(\lambda_c) = -1. \tag{15}$$

For $\lambda < \lambda_c$, we would like to show

$$\mathbb{E}\left[|h'_{t^2}(i)|^\lambda\right] \asymp t^{-\zeta(\lambda)}$$

and that the expectation is concentrated on an event for which

$$|h'_{t^2}(i)| \approx t^{-\zeta'(\lambda)}.$$

We will actually show a slightly weaker version (we will show the stronger version for a certain range of λ, see Section 9).

Let $I(t,m)$ denote the indicator function of the event

$$|X_t| \leq m\sqrt{t}, \qquad \frac{1}{m} \leq \frac{Y_t}{\sqrt{t}} \leq m.$$

Note that the "typical" values for X_t, Y_t are of order $\sqrt{t}$; in fact it is not difficult to show that

$$\lim_{m\to\infty} \mathbb{E}[I(t,m)] = 1.$$

For fixed m,

$$\mathbb{E}\left[|h_t'(i)|^\lambda\, I(t,m)\right] \asymp t^{-\zeta/2}\, \mathbb{E}\left[M_t\, I(t,m)\right] \leq t^{-\zeta/2}\, \mathbb{E}\left[M_t\right] = t^{-\zeta/2}.$$

(The implicit constants depend on m as well as ζ and κ.) One consequence of the next few sections is the following.

Proposition 4.1. *If $\lambda < \lambda_c$, there exist c, m such that*

$$\mathbb{E}[M_t\, I(m,t)] \geq c.$$

In particular, there exists c_1 such that

$$\mathbb{E}[|h_t'(i)|^\lambda] \geq \mathbb{E}[|h_t'(i)|^\lambda\, I(t,m)] \geq c_1\, t^{-\zeta/2}. \tag{16}$$

We will now explain why $\zeta^*(\lambda) \neq \zeta(\lambda)$ for $\lambda > \lambda_c$. Using the ideas of the previous section, we can show that for $\lambda < \lambda_c$ the expectation in $\mathbb{E}[|h_t'(i)|^\lambda\, I(t,m)]$ concentrates (roughly speaking) on an event on which

$$|h_{t^2}'(i)| \approx t^{-\zeta'(\lambda)}.$$

However, (6) tells us that

$$|h_{t^2}'(i)| \leq c\,t.$$

Hence, for $\lambda \geq \lambda_c$, the expectation is concentrated on an event with $|h_{t^2}'(i)| \asymp t$, and

$$\mathbb{E}[|h_{t^2}'(i)|^\lambda\, I(m,t)] \asymp t^{(\lambda-\lambda_c)-\zeta_c}, \quad \lambda \geq \lambda_c.$$

We can write

$$\mathbb{E}[|h_{t^2}'(i)|^\lambda\, I(m,t)] \asymp t^{-\zeta^*(\lambda)},$$

where

$$\zeta^*(\lambda) = \begin{cases} \zeta(\lambda) & \lambda \leq \lambda_c \\ \zeta(\lambda) + (\lambda_c - \lambda) & \lambda \geq \lambda_c. \end{cases}$$

The fact that the measure concentrates on (approximately) the same event for $\lambda \geq \lambda_c$ is reflected in the linearity of the function ζ^*.

Remark. While there is a "phase transition" in the expectation at $\lambda = \lambda_c$ there is no corresponding transition as $\lambda \to -\infty$. Using either the Beurling projection theorem (see, e.g., [5]) or (19) it can be seen that $|h_{t^2}'(i)| \geq c\,t^{-1}$. This value is obtained if U_t is constant and $Z_t(i)$ goes deterministically upward. Since $\zeta'(\lambda) < 1$ for all λ, the weighted measure never concentrates on these paths.

Example. When studying the Hausdorff dimension of an SLE path, one is led to study $|h'_{t^2}(i)|^d$, where

$$d = 1 + \frac{1}{4a} = 1 + \frac{\kappa}{8}.$$

For $\kappa < 8$, this turns out to be the Hausdorff dimension of the paths. Note that if $r = 1$, then $\lambda = d$. For $\kappa < 8$, we have $r < r_c, \lambda < \lambda_c$, and

$$\zeta^*(\lambda) = \zeta(\lambda) = 1 - \frac{1}{4a} = 2 - d.$$

Example. Second moment arguments for Hausdorff dimension lead to studying the $2d$-moment, i.e.,

$$\lambda = 2 + \frac{1}{2a}.$$

There are two regimes to consider.

- $5/4 \leq a < \infty$. In this range $2d \leq \lambda_c$. We have
$$r = r(2d) = 2a + 1 - 2a\sqrt{1 - \frac{1}{a} - \frac{1}{4a^2}},$$
$$\zeta^*(2d) = \zeta(2d) = 1 + \sqrt{1 - \frac{1}{a} - \frac{1}{4a^2}}.$$
- $1/4 < a \leq 5/4$. In this range
$$2d = \lambda_c + \left[1 + \frac{5}{16a} - a\right],$$
where the term in brackets is nonnegative, and hence
$$\zeta^*(2d) = \zeta(\lambda_c) - \left[1 + \frac{5}{16a} - a\right] = 2a - \frac{3}{8a} - 1.$$
Note that $\zeta^*(2d) = 0$ if $a = 3/4$ and $\zeta^*(2d) < 0$ for $1/4 < a < 3/4$.

5. Change of time

In this section we fix $z \in \mathbb{H}$ and write X_t, Y_t, Z_t, S_t for $X_t(z), Y_t(z), Z_t(z), S_t(z)$ although it is important to remember that these quantities depend on the starting point z. Since Y_t is differentiable and strictly increasing, we can find a new parametrization (depending on z) such that $\log Y_t$ grows linearly. To be more precise, let

$$\sigma(t) = \inf\{s : Y_s = e^{at}\}, \quad \hat{Y}_t = Y_{\sigma(t)} = e^{at}, \quad \hat{X}_t = X_{\sigma(t)}, \quad \hat{Z}_t = Z_{\sigma(t)},$$
$$K_t = e^{-at}\,\hat{X}_t, \quad \hat{S}_t = S_{\sigma(t)} = \frac{e^{at}}{|\hat{Z}_t|} = (K_t^2+1)^{-1/2}, \quad \hat{h}_t = h_{\sigma(t)}.$$

Lemma 5.1. *If* $z = x + e^{at_0} i$,

$$\partial_t \sigma(t) = |\hat{Z}_t|^2, \quad \sigma(t) = \int_{t_0}^t |\hat{Z}_s|^2\,ds = \int_{t_0}^t e^{2as}\,(K_s^2+1)\,ds. \tag{17}$$

Proof. Since $\partial_t \hat{Y}_t = a\,\hat{Y}_t$, (3) and the chain rule imply

$$a\,\hat{Y}_t = \frac{a\,\hat{Y}_t}{|\hat{Z}_t|^2}\,[\partial_t \sigma(t)]\,. \qquad \square$$

Using (3) we get

$$d\hat{X}_t = -a\,\hat{X}_t\,dt + |\hat{Z}_t|\,d\tilde{B}_t,$$
$$dK_t = -2a\,K_t\,dt + \sqrt{K_t^2+1}\,d\tilde{B}_t, \tag{18}$$

where $\tilde{B}_t$ denotes the standard Brownian motion

$$\tilde{B}_t = \int_0^{\sigma(t)} \frac{1}{|Z_s|}\,dB_s.$$

From (5) and (17), we see that

$$\partial_t|\hat{h}'_t(z)| = |h'_{\sigma(t)}(z)|\,|Z_t|^2 = |h'_{\sigma(t)}(z)|\,\frac{a\,(\hat{X}_t^2 - \hat{Y}_t^2)}{|\hat{Z}_t|^2} = a\,|\hat{h}'_t(z)|\,\left[1 - 2\,\hat{S}_t^2\right],$$

and hence,

$$|\hat{h}'_t(x+i)| = \exp\left\{a\int_0^t [1 - 2\,\hat{S}_s^2]\,ds\right\}.$$

Note that this implies

$$e^{-at} \le |\hat{h}'_t(x+i)| \le e^{at}. \tag{19}$$

6. The SDE (18)

Let

$$q = 2a + \frac{1}{2} - r, \tag{20}$$

and note that if $r < r_c$, then $q > 0$. We will study the equation (18) which we write as

$$dK_t = \left(\frac{1}{2} - q - r\right)\,K_t\,dt + \sqrt{K_t^2+1}\,d\tilde{B}_t. \tag{21}$$

For this section, we consider q, r as the given parameters, and we define a by (20). A simple application of Itô's formula gives the following lemma.

Lemma 6.1. *Suppose J_t satisfies*

$$dJ_t = -\,(q+r)\,\tanh J_t\,dt + d\tilde{B}_t. \tag{22}$$

Then $K_t = \sinh J_t$ satisfies (21).

Let

$$L_t = t - \int_0^t \frac{2\,ds}{K_s^2+1} = t - \int_0^t \frac{2\,ds}{\cosh^2 J_s}.$$

Note that $-t \le L_t \le t$ and

$$\partial_t[e^{L_t}] = e^{L_t}\,\left[1 - \frac{2}{\cosh^2 J_t}\right]. \tag{23}$$

As in (17), we let

$$\sigma(t) = \int_0^t e^{2as}[K_s^2 + 1]\, ds = \int_0^t e^{2as} \cosh^2 J_s \, ds \geq \int_0^t e^{2as}\, ds = \frac{1}{2a}\, \left[e^{2at} - 1\right]. \tag{24}$$

Although all the quantities above are defined in terms of J_t, it is useful to note that in the notation of the previous section,

$$|\hat{h}_t'(z)| = e^{aL_t}, \quad \hat{Y}_t = e^{at}, \quad \frac{\hat{X}_t}{\hat{Y}_t} = K_t, \quad [\hat{S}_t]^{-1} = \cosh J_t.$$

We let

$$N_t = e^{\nu L_t}\, e^{\xi t}\, [\cosh J_t]^r, \tag{25}$$

where

$$\nu = a\lambda = r\left(a + \frac{1}{2}\right) - \frac{r^2}{4} = r\left(\frac{q}{2} + \frac{1}{4}\right) + \frac{r^2}{4},$$
$$\xi = a\zeta = a\lambda - \frac{r}{2} = ar - \frac{r^2}{2} = r\left(\frac{q}{2} - \frac{1}{4}\right) + \frac{r^2}{4}.$$

In the notation of the previous section, $N_t = M_{\sigma(t)}$. We have written N_t and defined the exponents ν, ξ so they depend only on r, q and not on a. Note that

$$r(\nu) = -\left(q + \frac{1}{2}\right) + \sqrt{\left(q + \frac{1}{2}\right)^2 + 4\nu},$$

$$\xi(\nu) = \nu + \left(\frac{q}{2} + \frac{1}{4}\right) - \frac{1}{2}\sqrt{\left(q + \frac{1}{2}\right)^2 + 4\nu},$$

Since N_t is M_t sampled at an increasing family of stopping times, the next proposition is no surprise.

Proposition 6.2. *Suppose $r \in \mathbb{R}$ and ξ, ν are defined as in Proposition* 2.1. *Then N_t as defined in* (25) *is a positive martingale satisfying*

$$dN_t = N_t\, r\, [\tanh J_t]\, d\tilde{B}_t. \tag{26}$$

In particular,

$$\mathbb{E}^x[e^{\nu L_t}\, [\cosh J_t]^r] = e^{-\xi t}\, \mathbb{E}^x[N_t] = [\cosh x]^r\, e^{-\xi t}.$$

Proof. Itô's formula gives (26). If we use Girsanov's theorem, the weighted paths satisfy

$$dJ_t = -q\, [\tanh J_t]\, dt + dW_t, \tag{27}$$

where

$$W_t = \tilde{B}_t - r \int_0^t \tanh J_s\, ds,$$

is a standard Brownian motion in the new measure. Since $|\tanh| \leq 1$, it is straightforward to show that this equation does not have explosion in finite time, and hence we can see that N_t is actually a martingale. □

7. The SDE (27) for $q > 0$

We now focus our discussion on the equation (27) which has only one parameter q that we will assume is positive. Recall from (20) that this corresponds to $r < r_c$.

Lemma 7.1. *Suppose $q > 0$ and J_t satisfies* (27)*.*

- *J_t is a positive recurrent process with invariant density*
$$v_q(x) = \frac{C_q}{\cosh^{2q} x}, \qquad -\infty < x < \infty,$$
where
$$C_q = \frac{\Gamma(q+\frac{1}{2})}{\Gamma(\frac{1}{2})\,\Gamma(q)}.$$
Moreover,
$$\int_{-\infty}^{\infty} \left[1 - \frac{2}{\cosh^2 x}\right] v_q(x)\,dx = \mu := \frac{1-2q}{1+2q}.$$
- *If*
$$F(x) = F_q(x) = \int_0^x [\cosh y]^{2q}\,dy,$$
then $F(J_t)$ is a local martingale.
- *There exists $c = c_q$ such that for $0 \le y < x, k \ge 0$,*
$$\mathbb{P}^y\{|J_t| \ge x \textit{ for some } k \le t \le k+1\} \le c \left(\frac{\cosh y}{\cosh x}\right)^{2q}. \tag{28}$$

Proof. The computation of the invariant density is standard (but see the appendix for a derivation). Since
$$F''(x) = 2q\,[\tanh x]\,F'(x),$$
Itô's formula shows that $F(J_t)$ is a local martingale. Note that as $x \to \infty$, $F(x) \sim (2q)^{-1}\,[\cosh x]^{2q}$.

It suffices to prove (28) for $x \ge 1$. Assume first that $y = 0$. A coupling argument shows that the probability in (28) is bounded above by the corresponding probability where J_0 has the law of the invariant distribution. Suppose J_0 has density v_q, and let
$$Y = Y_{k,x} = \int_k^{k+2} 1\{|J_t| \ge x-1\}\,dt.$$
Then
$$\mathbb{E}[Y] = 2\int_{|y|\ge x-1} v_q(y)\,dy \le \frac{c}{\cosh^{2q} x}.$$
Since the drift in (27) is bounded, we can see from the strong Markov property that for some $\delta > 0$,
$$\mathbb{E}[Y \mid |J_t| \ge x \text{ for some } k \le t \le k+1] \ge \delta,$$
and hence
$$\mathbb{P}\{|J_t| \ge x \text{ for some } k \le t \le k+1\} \le \delta^{-1}\,\mathbb{E}[Y].$$

If $0 < y < x$, let $T = T(0,x)$ be the first time that $J_t \in \{0,x\}$. Using the strong Markov property we see that

$$\mathbb{P}^y\{|J_t| \geq x \text{ for some } k \leq t \leq k+1 \;;\; J_T = 0\} \leq \frac{c}{[\cosh x]^{2q}}.$$

Hence,

$$\mathbb{P}^y\{|J_t| \geq x \text{ for some } k \leq t \leq k+1\} \leq \frac{c}{[\cosh x]^{2q}} + \mathbb{P}^y\{J_T = x\}.$$

Applying the optional sampling theorem to the martingale $F(J_{T\wedge t})$, we see that

$$\mathbb{P}^y\{J_T = x\} = \frac{F(y)}{F(x)} \leq c\left(\frac{\cosh y}{\cosh x}\right)^{2q}. \qquad \square$$

Remark. For $y = 0$, the estimate (28) is sharp for large t. For $y > 0$, the estimate is not sharp for large t; in fact, for large t one gets the same estimate as for $y = 0$. However, how large t needs to be depends on y, x and (28) is the best one can do if one wants a uniform estimate for all t, y, x.

As in the previous section, we let

$$L_t = \int_0^t \left[1 - \frac{2}{\cosh^2 J_s}\right] ds.$$

It follows from the previous lemma that as $t \to \infty$,

$$t^{-1}\,\mathbb{E}[L_t] \sim \int_{-\infty}^{\infty} \left[1 - \frac{2}{\cosh^2 x}\right] v_q(x)\,dx = \mu = \frac{1-2q}{1+2q}.$$

In fact, we claim that

$$L_t = \mu\,t + O(t^{1/2}).$$

In Section 10 we will need large (or, as sometimes called, moderate) deviations, i.e., probabilities that $|L_t - \mu t|$ is much larger than $t^{1/2}$. Let

$$\overline{L}_t = L_t - \mu t.$$

The standard way to obtain large deviation results is to obtain a bound on an exponential moment. The martingales allow us to do that rather easily here. The value b in the next proposition is not special; in fact, a small modification of the proof shows that the expectation is bounded for all b. However, we only need to use one value, and the proof simplifies slightly by restricting to this case. This proposition should be compared to Proposition 3.1.

Proposition 7.2. *Suppose J_t satisfies* (27) *with $q > 0$. Then there exists $c < \infty$ such that for all $0 \leq s < t$,*

$$\mathbb{E}\left[\exp\left\{\frac{b\,|\overline{L}_t - \overline{L}_s|}{\sqrt{t-s}}\right\}\right] \leq c, \qquad (29)$$

where $b = 2q + 1$.

Proof. Since $|\overline{L}_t - \overline{L}_s| \leq 2(t-s)$, it suffices to prove the bound for $t-s$ sufficiently large. For this proof, we will assume that $t-s \geq (4/q)^2$, i.e., $4/\sqrt{t-s} \leq q$.

We will first show that there is a c such that

$$\mathbb{E}\left[\exp\left\{\pm b\,\frac{(\overline{L}_t - \overline{L}_s)}{\sqrt{t-s}}\right\}\,[\cosh J_t]^{\pm\frac{4}{\sqrt{t-s}}}\right] \leq c. \tag{30}$$

For $\beta \in \mathbb{R}$, let

$$\delta = \delta(\beta) = \beta\left(\frac{q}{2} + \frac{1}{4}\right) - \frac{\beta^2}{4}, \qquad \rho = \beta\left(\frac{q}{2} - \frac{1}{4}\right) - \frac{\beta^2}{4}.$$

Using Proposition 6.2 with (q,r) replaced with $(q-\beta,\beta)$, we see that

$$O_t := e^{\delta L_t}\, e^{\rho t}\,[\cosh J_t]^\beta = e^{\delta \overline{L}_t}\, e^{(\rho+\delta\mu)t}\,[\cosh J_t]^\beta,$$

is a martingale. In particular, for $s < t$,

$$\mathbb{E}\left[e^{\delta(\overline{L}_t - \overline{L}_s)}\, e^{(\rho+\delta\mu)(t-s)}\,(\cosh J_t)^\beta\right] = \mathbb{E}\left[(\cosh J_s)^\beta\right]. \tag{31}$$

If $\beta \leq q$, then we can use (28) to see there is a $c < \infty$ such that

$$\mathbb{E}\left[(\cosh J_s)^\beta\right] < c.$$

If we apply this bound with $\beta = \pm 4/\sqrt{t-s} \leq q$, then

$$\delta = \pm\frac{2q+1}{\sqrt{t-s}} + \frac{1}{4(t-s)},$$

$$\rho = \pm\frac{2q-1}{\sqrt{t-s}} - \frac{1}{4(t-s)} = -\mu\delta + O\left(\frac{1}{t-s}\right).$$

Hence,

$$\exp\left\{\pm b\,\frac{\overline{L}_t - \overline{L}_s}{\sqrt{t-s}}\right\} \leq c\, e^{\delta(\overline{L}_t - \overline{L}_s)}\, e^{(\rho+\delta\mu)(t-s)},$$

and (30) follows from (31).

Clearly, (30) implies

$$\mathbb{E}\left[\exp\left\{b\,\frac{\overline{L}_t - \overline{L}_s}{\sqrt{t-s}}\right\}\right] < \infty.$$

To prove (29), we also need to show the corresponding inequality with b replaced with $-b$. By choosing y sufficiently large, we can see from (28) that

$$\mathbb{P}\{|J_t| \geq y\sqrt{t-s}\} \leq e^{-2b\sqrt{t-s}}.$$

Since $|\overline{L}_t - \overline{L}_s| \leq 2\sqrt{t-s}$, this implies

$$\mathbb{E}\left[\exp\left\{-b\,\frac{\overline{L}_t - \overline{L}_s}{\sqrt{t-s}}\right\}\;;\;|J_t| \geq y\sqrt{t}\right] \leq c.$$

However, if $|J_t| \leq y\sqrt{t}$, then $[\cosh J_t]^{-4/\sqrt{t}}$ is bounded below and hence (30) implies

$$\mathbb{E}\left[\exp\left\{-b\,\frac{\overline{L}_t - \overline{L}_s}{\sqrt{t-s}}\right\}\;;\;|J_t| \leq y\sqrt{t}\right] \leq c. \qquad \square$$

The next proposition makes precise the idea that $J_s = O(1)$ and

$$L_s = \mu\, s + O(\sqrt{s}), \quad L_t - L_s = \mu\,(t-s) + O(\sqrt{t-s}).$$

It is phrased in a way that is used in Section 10. In particular, it considers an event where the error is of order a constant times $\sqrt{s}$ or $\sqrt{t-s}$ when s is small but allows a somewhat larger error for other values of s. Let

$$F(s,t) = 2 + \min\{s, t-s\}.$$

Proposition 7.3. *Suppose J_t satisfies* (27) *with $q > 0$. For each $u, t > 0$, let $E_{t,u}$ be the event that the following holds for all $0 \leq s \leq t$:*

$$|J_s| \leq u \, \log F(s,t),$$

$$|\overline{L}_s| \leq u\,\sqrt{s}\, \log(s+2),$$

$$|\overline{L}_t - \overline{L}_s| \leq u\,\sqrt{t-s}\, \log(t-s+2),$$

Then

$$\lim_{u\to\infty} \inf_{t>0} \mathbb{P}(E_{t,u}) = 1.$$

Remark. For future reference we note that on the event $E_{u,t}$,

$$\cosh^2 J_s \leq e^{2|J_s|} \leq F(s,t)^{2u}.$$

In particular, there exists C_u such that for all $t_1 < t$,

$$\sigma(t_1) := \int_0^{t_1} e^{2as}\,[\cosh^2 J_s]\, ds \leq C_u\, F(t_1,t)^u\, e^{2at_1}. \tag{32}$$

Proof. For ease we assume that t is a positive integer. From (28), we see that there is a c_1 such that

$$\begin{aligned} \mathbb{P}\{|J_s| \geq u\, \log(s+2) \text{ for some } k \leq s \leq k+1\} & \\ \leq \quad & \mathbb{P}\{|J_s| \geq u\, \log(k+2) \text{ for some } k \leq s \leq k+1\} \\ \leq \quad & c_1\,(k+2)^{-2qu}. \end{aligned}$$

Therefore,

$$\mathbb{P}\{|J_s| \geq u\, \log(s+2) \text{ for some } s \geq 0\} \leq c_1 \sum_{k=0}^{\infty} (k+2)^{-2qu}, \tag{33}$$

and the right-hand side goes to zero as $u \to \infty$. A similar argument shows that

$$\mathbb{P}\{|J_s| \geq u\, \log(t-s+2) \text{ for some } 0 \leq s \leq t\} \leq c_1 \sum_{k=0}^{\infty} (k+2)^{-2qu}. \tag{34}$$

Since $|\overline{L}_t - \overline{L}_s| \leq 2(t-s)$,

$$\mathbb{P}\{|\overline{L}_s| \geq u\,\sqrt{s}\, \log(s+2) \text{ for some } k \leq s \leq k+1\} \leq$$
$$\mathbb{P}\{|\overline{L}_k| \geq u\,\sqrt{k}\, \log(k+2) - 2\}.$$

Using (29) and Chebyshev's inequality, we see that the right-hand side is bounded by a constant times $(k+2)^{-bu}$. Similarly,

$$\mathbb{P}\{|\overline{L}_t - \overline{L}_s| \geq u\,\sqrt{t-s}\,\log(t-s+2) \text{ for some } k \leq s \leq k+1\} \leq c\,(k+2)^{-bu}.$$

The argument proceeds as in (33) and (34). □

8. Lower bound

Having analyzed the one-variable equation for J_t we return to the original problem. Let $r < 2a + \frac{1}{2}$ and $q = 2a + \frac{1}{2} - r > 0$. We fix a u as in Proposition 7.3 such that for all t,

$$\mathbb{P}_*(E_{t,u}) \geq \frac{1}{2},$$

and we write just E_t for $E_{t,u}$. Here we write $\mathbb{P}_*$ for the probability measure to distinguish it from the probability measure under which the expectation $\mathbb{E}$ below is defined. Let $C = C_u$ be a constant such that (32) holds. We fix u and allow all constants in this section to depend on u. We let $z = i$ and write $X_t, Y_t, M_t, \ldots$ for $X_t(i), Y_t(i), M_t(i), \ldots$

Let M_t be is the martingale associated to r. which we can write as M_t as

$$M_t = |h_t'(i)|^\lambda\, Y_t^\zeta\, [\cosh \tilde{J}_t]^r,$$

where $\tilde{J}_t$ is defined by

$$\sinh \tilde{J}_t = \frac{X_t}{Y_t}.$$

Note that $J_t = \tilde{J}_{\sigma(t)}$. Then Proposition 7.3 can be rewritten as

$$\begin{aligned} 1 &\geq \mathbb{E}\left[|h_{\sigma(t)}'(i)|^\lambda\, Y_{\sigma(t)}^\zeta\, (\cosh \tilde{J}_{\sigma(t)})^r\, 1_{E_t}\right] \\ &= e^{a\zeta t}\mathbb{E}\left[|h_{\sigma(t)}'(i)|^\lambda\, (\cosh \tilde{J}_{\sigma(t)})^r\, 1_{E_t}\right] \geq \frac{1}{2}. \end{aligned}$$

On the event E_u, J_t is uniformly bounded. Therefore, this implies

$$\mathbb{E}\left[|h_{\sigma(t)}'(i)|^\lambda\, 1_{E_t}\right] \asymp e^{-a\zeta t}.$$

We now derive some bounds that hold on the event E_t. Recall that

$$|h_{\sigma(t)}'(i)| = e^{aL_t} = e^{at\mu}\, e^{a\overline{L}_t}.$$

The proposition implies for all $0 \leq s \leq t$,

$$\cosh J_s \leq F(s,t)^u,$$

$$\exp\left\{-u\sqrt{s}\log(s+2)\right\} \leq e^{-as\mu}\,|h_{\sigma(s)}'(i)| \leq \exp\left\{u\sqrt{s}\log(s+2)\right\},$$

$$\begin{aligned} \exp\left\{-u\sqrt{t-s}\log(t-s+2)\right\} &\leq e^{-a(t-s)\mu}\,\frac{|h_{\sigma(t)}'(i)|}{|h_{\sigma(s)}'(i)|} \\ &\leq \exp\left\{u\sqrt{t-s}\log(t-s+2)\right\}. \end{aligned}$$

The Loewner equation implies that

$$\sigma(s) \geq \frac{e^{2as}-1}{2a},$$

and the proposition gives the upper bound

$$\sigma(s) \leq \int_0^s e^{2av} \cosh^2 J_v \, dv \leq \int_0^s e^{2av} \, (v+2)^u \, J_v \, dv \leq c \, (s+2)^{2u} \, e^{2as},$$

$$\sigma(s) \leq \int_0^s e^{2av} \cosh^2 J_v \, dv \leq \int_0^s e^{2av} \, (t-v+2)^{2u} \, J_v \, dv \leq c \, (t-s+2)^u \, e^{2as},$$

i.e.,

$$\sigma(s) \leq c \, F(s,t)^{2u} \, e^{2as}.$$

By inverting this, we get

$$s - c \min\{\log(s+2), \log(t-s+2)\} \leq \frac{1}{a} \log Y_{e^{2as}} \leq s + c, \quad 0 \leq s \leq t.$$

This yields

$$\cosh \tilde{J}_{e^{2as}} \leq c \, F(s,t)^u$$

$$\exp\left\{-u\sqrt{s} \log(s+2)\right\} \leq e^{-as\mu} \, |h'_{e^{2as}}(i)| \leq \exp\left\{u\sqrt{s}\log(s+2)\right\},$$

$$\begin{aligned} \exp\left\{-u\sqrt{t-s}\log(t-s+2)\right\} \quad &\leq \quad e^{-a(t-s)\mu} \, \frac{|h'_{\sigma(t)}(i)|}{|h'_{e^{2as}}(i)|} \\ &\leq \quad \exp\left\{u\sqrt{t-s}\log(t-s+2)\right\}. \end{aligned}$$

Once we have this, we can continue the process from time $\sigma(t)$ to time ce^{2at}. From this, one can deduce the following which is used in Section 10. The statement is rather cumbersome, but it essentially follows from what we have done.

Theorem 8.1. *Suppose* $r < 2a + \frac{1}{2}$ *and* λ, ζ *are defined as in Proposition* 2.1. *Let*

$$\mu = \frac{1-2q}{1+2q} = \frac{2(r-2a)}{1+2a-r}, \quad F(s,t) = 2 + \min\{s, t-s\}.$$

For each b, u, *let* $A(b,u,t)$ *denote the indicator function of the event that the following holds for* $0 \leq s \leq t$:

$$b^{-1} \, \frac{e^{2as}}{F(s,t)^u} \leq Y_{e^{2as}} \leq b \, e^{as},$$

$$\cosh \tilde{J}_{e^{2as}} \leq b \, F(s,t)^u,$$

$$b^{-1} \exp\left\{-u\sqrt{s}\log(s+2)\right\} \leq e^{-as\mu} \, |h'_{e^{2as}}(i)| \leq b \exp\left\{u\sqrt{s}\log(s+2)\right\},$$

$$\begin{aligned} b^{-1} \exp\left\{-u\sqrt{t-s}\log(t-s+2)\right\} \, |h'_{e^{2as}}(i)| &\leq \\ e^{-a(t-s)\mu} \, |h'_{e^{2at}}(i)| \leq b \exp\left\{u\sqrt{t-s}\log(t-s+2)\right\} \, &|h'_{e^{2as}}(i)|. \end{aligned}$$

Then there exist b, u *such that for all* $t > 0$,

$$b^{-1} \leq \mathbb{E}\left[|h'_{e^{2at}}(i)|^\lambda \, A(b,u,t)\right] \leq b.$$

9. An upper bound

Here we prove a theorem which gives an upper bound on some of the moments of $|h'(z)|$ for a range of λ. The dependence on x is probably not optimal. Indeed, as remarked after Proposition 7.1, the estimates for large x used in the proof are not optimal.

Theorem 9.1. *Suppose*

$$0 < r < 6a - 2\sqrt{5a^2 - a}, \quad a \geq \frac{1}{4}, \tag{35}$$

$$0 < r < 2a + \frac{1}{2}, \quad a < \frac{1}{4}, \tag{36}$$

λ, ζ *are defined as in Proposition* 2.1 *and* $q = 2a + \frac{1}{2} - r$. *Then there exists a* $c < \infty$ *such that for all* $x \in \mathbb{R}$, $y > 0$,

$$\mathbb{E}[|h'_{(sy)^2}(xy + iy)|^\lambda] = \mathbb{E}[|h'_{s^2}(x+i)|^\lambda] \leq c\,(x^2+1)^{\frac{r+\zeta}{2}}\,[\log(x^2+2)]^{2q}\,(s+1)^{-\zeta}.$$

In particular, if $a > 1/4$, *there exists a* $c < \infty$, *such that for all* $x \in \mathbb{R}$,

$$\mathbb{E}[|h'_{s^2}(x+i)|^d] \leq c\,(x^2+1)^{1-\frac{1}{8a}}\,[\log(x^2+2)]^{4a-1}\,(s+1)^{d-2}.$$

The final assertion follows from the previous one by plugging in $r = 1$ which satisfies (35) for $a > \frac{1}{4}$.

Proof. By scaling we may assume $y = 1$ and without loss of generality, we assume $x = e^{al} \geq 0$. If $s \leq 1$, the Loewner equation implies that $|h'_{s^2}(x+i)| \asymp 1$, so we will assume $s = e^{at} \geq 1$. We write $X_s, Y_s, \ldots$ for $X_s(x+i), Y_s(x+i), \ldots$.

Consider the martingale

$$M_s = M_{s,r}(x+i) = |h'_s(x+i)|^\lambda\, Y_s^\zeta\, S_s^{-r}.$$

The conditions (35) and (36) imply that (36) holds for all a; $q > 0$; $r < 4a$; and

$$-2q < \zeta - 2q = r - \frac{r^2}{4a} - 2q < 0. \tag{37}$$

Recall that

$$\sigma(s) = \inf\{u : Y_u = e^{as}\},$$

and let $\hat{M}_s = M_{\sigma(s)}$. Let $\tau = \tau_t$ be the minimum of t and the smallest s such that

$$\hat{S}_s \leq (t - s + 1)\, e^{-a(t-s)}.$$

Let $\rho = \sigma(\tau)$ so that $S_\rho = \hat{S}_\tau$. (Note that τ is the time in the new parametrizaton, and $\rho = \sigma(\tau)$ is the corresponding amount of time in the original parameterization. If one considers curves modulo reparametrization, then ρ and τ represent the same stopping "time".) Note that

$$\rho = \sigma(\tau) = \int_0^\tau e^{2as}\, \hat{S}_s^{-2}\, ds \leq \int_0^t \frac{e^{2at}}{(t-s+1)^2}\, ds \leq e^{2at}.$$

For positive integer k, let $A_k = A_{k,t}$ be the event $\{t-k < \tau \leq t-k+1\}$. Since M_t is a martingale, $\tau \leq e^{2at}$, and the event A_k depends only on $M_s, 0 \leq s \leq \tau$, the optional sampling theorem gives

$$\mathbb{E}[M_{e^{2at}}\, 1_{A_k}] = \mathbb{E}[M_\rho\, 1_{A_k}] = \mathbb{E}[\hat{M}_\tau\, 1_{A_k}].$$

Since Y_t increases with t, we know that on the event A_k,

$$Y_{e^{2at}} \geq Y_\rho \geq e^{at}\, e^{-ak}, \qquad S_\rho^2 \asymp e^{-2ak}\, k^2.$$

The Girsanov theorem implies that

$$\mathbb{E}[\hat{M}_\tau\, 1_{A_k}] = M_0\, \mathbf{Q}(A_k) \leq c\, e^{alr}\, \mathbf{Q}(A_k),$$

where $\mathbf{Q}$ denotes the measure obtained by weighting by the martingale $\hat{M}$. From (28) we know that

$$\mathbf{Q}(A_k) \leq c\, e^{2aq(l-k)}\, k^{2q}.$$

Therefore, if we write $s = e^{at}$,

$$\begin{aligned} s^\zeta\, \mathbb{E}\left[|h'_{s^2}(x+i)|^\lambda\, 1_{A_k}\right] &\leq c\, e^{ak\zeta}\, \mathbb{E}\left[|h'_{s^2}(x+i)|^\lambda\, Y_{s^2}\, 1_{A_k}\right] \\ &\leq c\, e^{ak\zeta}\, \mathbb{E}[M_{s^2}\, 1_{A_k}] \\ &= c\, e^{ak\zeta}\, \mathbb{E}[M_\rho\, 1_{A_k}] \\ &\leq c\, e^{ak\zeta}\, e^{alr}\, \mathbf{Q}(A_k). \end{aligned}$$

Therefore,

$$s^\zeta\, \mathbb{E}\left[|h'_{s^2}(x+i)|^\lambda\right] \leq c\, e^{arl} \left[\sum_{k\leq l} e^{ak\zeta}\, \mathbf{Q}(A_k) + e^{al\zeta} \sum_{k>l} k^{2q}\, e^{a(k-l)(\zeta-2q)}\right].$$

Using (37), we can sum over k to get

$$s^\zeta\, \mathbb{E}\left[|h'_{s^2}(x+i)|^\lambda\right] \leq c\, e^{arl}\, l^{2q}\, e^{a\zeta l}. \qquad \square$$

10. Hausdorff dimension

We will prove that for $\kappa < 8$, the Hausdorff dimension of the paths is $d = 1 + \frac{\kappa}{8}$. We will only prove the lower bound which is the hard direction; the upper bound was proved by Rohde and Schramm [10] and we sketch the proof in the next paragraph. Since Hausdorff dimension is preserved under conformal maps, it is easy to use the independence of the increments of Brownian motion to conclude that there is a d_* such that with probability one $\dim_h[\gamma[t_1, t_2]] = d_*$ for all $t_1 < t_2$. Using this and the upper bound, we can see that it suffices to prove that for all $\alpha < d$,

$$\mathbb{P}\{\dim_h(\gamma[1,2]) \geq \alpha\} > 0. \tag{38}$$

The computation (and rigorous upper bound) of the dimension was done by Rohde and Schramm who first noted that

$$M_t = M_t(z) = \Upsilon_t^{d-2}\, S_t^{4a-1}$$

is a local martingale, where

$$\Upsilon_t = \frac{\mathrm{Im} g_t(z)}{|g_t'(z)|}, \qquad S_t = [\sin \arg(g_t(z) - V_t)].$$

The Koebe (1/4)-theorem (see Section 10.4) shows that

$$\frac{1}{4}\,\Upsilon_t \leq \mathrm{dist}\,(0, \gamma[0,t] \cap \mathbb{R}) \leq 4\,\Upsilon_t.$$

If T_ϵ is the first time that $\Upsilon_t \leq \epsilon$, then the optional sampling theorem can be used to see that

$$M_0(z) = \mathbb{E}[M_{T_\epsilon}; T_\epsilon < \infty] = \epsilon^{d-2}\,\mathbb{E}[S_{T_\epsilon}^{4a-1}; T_\epsilon < \infty].$$

By using the Girsanov theorem [6], one can show that there is a c_* such that

$$\mathbb{E}[S_{T_\epsilon}^{4a-1}; T_\epsilon < \infty] \sim c_*\,\mathbb{P}\{T_\epsilon < \infty\}.$$

which shows that

$$\mathbb{P}\{\Upsilon_\infty \leq \epsilon\} \sim c_*^{-1}\,\epsilon^{2-d}\,M_0(z).$$

In particular,

$$\mathbb{P}\{\mathrm{dist}[\gamma(0,\infty), z] \leq \epsilon\} \asymp M_0(z)\,\epsilon^{2-d}.$$

From this the upper bound for the dimension follows easily.

The lower bound follows from standard techniques provided that one has a "two-point" estmate

$$\mathbb{P}\{\mathrm{dist}[\gamma(0,\infty), z] \leq \epsilon, \mathrm{dist}[\gamma(0,\infty), w] \geq \epsilon\} \asymp \epsilon^{2-d}\,\left(\frac{|z-w|}{\epsilon}\right)^{2-d}.$$

This was successfully established by Beffara [1] although the argument is somewhat complicated.

We take a different approach to proving the lower bound by using the reverse Loewner flow. As in Beffara's approach, we construct a measure on the curve that is in some sense a d-dimensional measure and use a version of Frostman's lemma.

10.1. A version of Frostman's lemma

The main tool for proving lower bounds for Hausdorff dimension is Frostman's lemma (see [3, Theorem 4.13]), a version of which we recall here: if $A \subset \mathbb{R}^m$ is compact and μ is a Borel measure with $\mu(\mathbb{R}^m \setminus A) = 0$, $\mu(A) > 0$, and

$$\mathcal{E}_\alpha(\mu) := \int\int \frac{\mu(dx)\,\mu(dy)}{|x-y|^\alpha} < \infty, \tag{39}$$

then the Hausdorff-α measure of A is infinite. In particular, $\dim_h(A) \geq \alpha$. The following two lemmas summarize a standard technique for proving lower bounds of dimensions of random sets.

Lemma 10.1. *Suppose $A \subset \mathbb{R}^m$ is compact and $0 < d \leq m$. Suppose $c > 0$ and $r : (0,d) \to (0,\infty)$. Suppose there exists a decreasing sequence of compact sets A_n with $\cap_n A_n = A$ and a sequence of Borel measures μ_n with $\mu_n(\mathbb{R}^m \setminus A_n) = 0$, $\mu(A_n) \geq c$, and $\mathcal{E}_\alpha(\mu_n) \leq r(\alpha)$ for $0 < \alpha < d$. Then $\dim_h(A) \geq d$.*

Proof. (sketch) There exists a subsequence μ_{n_k} such that μ_{n_k} converges to a measure μ. One needs only check that $\mu(\mathbb{R}^m \setminus A) = 0$, $\mu(A) \geq c$ and $\mathcal{E}_\alpha(\mu) \leq r(\alpha)$. □

Lemma 10.2. *Suppose $A \subset \mathbb{R}^m$ is a random compact subset of $\mathbb{R}^m$ and $0 < d \leq m$. Suppose $0 < c_1 < c_2 < \infty$, $r : (0,d) \to (0,\infty)$, and $\delta_n \to 0$. Let $A_n = \{x \in \mathbb{R}^m : \operatorname{dist}(x,A) \leq \delta_n\}$. Suppose there exists a sequence of random Borel measures μ_n and a sequence $\delta_n \to 0$ such that the following is true for each n and each $\alpha < d$:*

$$\begin{aligned} \mathbb{E}\left[\mu_n(A_n)\right] &\geq c_1, \\ \mathbb{E}\left[\mu_n(A_n)^2\right] &\leq c_2, \\ \mathbb{E}\left[\mathcal{E}_\alpha(\mu_n)\right] &\leq r(\alpha), \\ \mu_n\left(\mathbb{R}^m \setminus A_n\right) &= 0. \end{aligned}$$

Then,

$$\mathbb{P}\left\{\dim_h(A) \geq d\right\} \geq \frac{c_1^2}{c_2}.$$

Proof. Let $q = c_1^2/c_2$. Standard “second moment” arguments (see, e.g., [5, Lemma A.15]) show that the first two inequalities imply that for every $p > 0$ there is an $\epsilon > 0$ such that

$$\mathbb{P}\{\mu_n(A_n) \geq \epsilon\} \geq q - p.$$

Since

$$\mathbb{P}\{\mathcal{E}_\alpha(\mu_n) \geq p^{-1}\, r(\alpha)\} \leq \frac{\mathbb{E}[\mathcal{E}_\alpha(\mu_n)]}{p^{-1}\, r(\alpha)} \leq p,$$

it follows that for each n,

$$\mathbb{P}\{\mu_n(A_n) \geq \epsilon, \mathcal{E}_\alpha(\mu_n) \leq p^{-1}\, r(\alpha)\} \geq q - 2p,$$

and hence

$$\mathbb{P}\{\mu_n(A_n) \geq \epsilon, \mathcal{E}_\alpha(\mu_n) \leq p^{-1}\, r(\alpha) \text{ infinitely often}\} \geq q - 2p.$$

On the event on the left-hand side, we have $\dim_h(A) \geq \alpha$ by the previous lemma. Therefore,

$$\mathbb{P}\{\dim_h(A) \geq \alpha\} \geq q - 2p.$$

Since this holds for all $\alpha < d$ and $p > 0$, the result follows. □

The next lemma is similar to many that have appeared before (see [5, A.3]), but the specific formulation may be new. For example, the assumptions (41) and (42) include subpower functions and are not quite as strong as if the functions were replaced by constants.

Definition. We will call a function $\phi : [0,\infty) \to (0,\infty)$ a *subpower function* if it is increasing, continuous, and

$$\lim_{x\to\infty} \frac{\log \phi(x)}{\log x} = 0,$$

i.e., ϕ grows slower than x^q for all $q > 0$.

Throughout this paper we will use ϕ, ψ to denote subpower functions. Similarly to the way arbitrary constants are handled, we will allow the particular value of the function to vary from line to line. We will not try to find the optimal subpower function for the results in this paper. We will use the fact that if ϕ, ψ are subpower functions, so are $\phi + \psi, \phi\psi, \phi \wedge \psi, \phi^k$.

Lemma 10.3. *Suppose $\eta : [0,1] \to \mathbb{R}^m$ is a random curve and*

$$\{F(j,n) : n = 1, 2, \ldots, j = 1, 2, \ldots, n\}$$

are nonnegative random variables all defined on the same probability space. Suppose $1 < d \leq m$, and there exist a subpower function ψ, $0 < \xi < 1$, and $c < \infty$ such that the following holds for $n = 1, 2, \ldots,$ and $1 \leq j \leq k \leq n$:

$$c^{-1} \leq \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[F(j,n)] \leq c, \tag{40}$$

$$\mathbb{E}[F(j,n)F(k,n)] \leq \left(\frac{n}{k-j+1}\right)^{\xi} \psi\left(\frac{n}{k-j+1}\right), \tag{41}$$

and

$$\left|\eta\left(\frac{j}{n}\right) - \eta\left(\frac{k}{n}\right)\right| \geq \left(\frac{k-j}{n}\right)^{\frac{1-\xi}{d}} \psi\left(\frac{n}{|j-k|+1}\right)^{-1} 1\{F(j,n)F(k,n) > 0\}. \tag{42}$$

Then

$$\mathbb{P}\{\dim_h(\eta[0,1]) \geq d\} > 0.$$

Remark. The proof constructs a measure supported on the curve. The nth approximation is a sum of measures $\mu_{j,n}$ which are multiples of Lebesgue measure on small discs centered at $\eta(j/n)$. The multiple at $\eta(j/n)$ is chosen so that the total mass $\mu_{j,n}$ is $F(j,n)/n$. In particular, if $F(j,n) = 0$, $\mu_{j,n}$ is the zero measure. To bound $\mathbb{E}\left[\mathcal{E}_\alpha(\mu_n)\right]$ we need to show that the measure is sufficiently spread out and (42) gives the necessary assumption. Note the assumption requires the inequality to hold only when $F(j,n)F(k,n) > 0$. The assumption implies that if $j < k$ and $\eta(j/n) = \eta(k/n)$ (or are very close), then at most one of $\mu_{j,n}$ and $\mu_{k,n}$ is nonzero.

Proof. We fix ξ, ψ, d and constants in this proof depend on ξ, ψ, m, d. Let $\mu_{j,n}$ denote the (random) measure that is a multiple of Lebesgue measure on the disk of radius $r_n := n^{\frac{\xi-1}{d}} \psi(n)/4$ about $\eta(j/n)$ where the multiple is chosen so that $\|\mu_{j,n}\| = n^{-1} F(j,n)$. Here $\|\cdot\|$ denotes total mass. Let $\nu_n = \sum_{j=1}^{n} \mu_{j,n}$. From (40), we see that

$$\mathbb{E}\left[\|\nu_n\|\right] \geq c_1,$$

and from (41) we see that

$$\begin{aligned}\mathbb{E}\left[\|\nu_n\|^2\right] &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}[F(j,n)\,F(k,n)]\\ &\leq \frac{2}{n}\sum_{k=1}^{n}\left(\frac{n}{k}\right)^{\xi}\,\psi\left(\frac{n}{k}\right)\\ &\leq 2\int_0^1 \frac{\psi(1/x)\,dx}{x^{\xi}} < \infty.\end{aligned}$$

The last inequality uses $\xi < 1$.

We will now show that for each $\alpha < d \leq m$, there is a C_α such that

$$\mathbb{E}[\mathcal{E}_\alpha(\nu_n)] = \sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\left[\int\int \frac{\mu_{j,n}(dx)\,\mu_{k,n}(dy)}{|x-y|^{\alpha}}\right] \leq C_\alpha. \qquad (43)$$

We will use the estimate

$$r^{-2m}\int_{|x-x_0|\leq r}\int_{|y-y_0|\leq r}\frac{d^m x\,d^m y}{|x-y|^{\alpha}} \leq c_\alpha \min\{r^{-\alpha}, |x_0-y_0|^{-\alpha}\}.$$

To estimate the terms with $j = k$, note that (41) with $j = k$ gives

$$\begin{aligned}\mathbb{E}\left[\int\int \frac{\mu_{j,n}(dx)\,\mu_{j,n}(dy)}{|x-y|^{\alpha}}\right] &\leq c\,\frac{\mathbb{E}[F(j,n)^2]}{n^2}\,r_n^{-\alpha}\\ &\leq c\,\frac{n^{\xi}\,\psi(n)}{n^2}\,n^{\frac{(1-\xi)\alpha}{d}}\,\psi(n)^{\alpha} = o(n^{-1})\end{aligned}$$

The last inequality uses $\alpha < d$. Therefore,

$$\mathbb{E}\left[\sum_{j=1}^{n}\int\int\frac{\mu_{j,n}(dx)\,\mu_{j,n}(dy)}{|x-y|^{\alpha}}\right] = o(1).$$

For $j < k$, we use the estimate

$$\int\int\frac{\mu_{j,n}(dx)\,\mu_{k,n}(dy)}{|x-y|^{\alpha}} \leq c\,\frac{F(j,n)\,F(k,n)}{n^2\,|\eta(j/n)-\eta(k/n)|^{\alpha}}.$$

Note that (40) and (41) combine to show that for each $\alpha < d$ there exist $c < \infty, \delta > 0$ (depending on α) such that

$$\mathbb{E}\left[\frac{F(j,n)\,F(k,n)}{|\eta(j/n)-\eta(k/n)|^{\alpha}}\right] \leq c\left(\frac{n}{k-j}\right)^{1-\delta}, \quad j < k. \qquad (44)$$

Combining this with (44) gives,

$$\sum_{1\leq j<k\leq n}\mathbb{E}\left[\int\int\frac{\mu_{j,n}(dx)\,\mu_{k,n}(dy)}{|x-y|^{\alpha}}\right] \leq c\,\frac{C_3}{C_4\,n^2}\sum_{1\leq j<k\leq n}\left(\frac{n}{k-j}\right)^{1-\delta} \leq C_\alpha.$$

This gives (43). The lemma now follows from Lemma 10.2. □

Remark. The usual form of this lemma chooses η to be the identity function and $d = 1-\xi$ in which case (42) is immediate. The condition (41) is often replaced with a stronger assumption where the subpower function ψ replaced with a constant.

It will be useful for us to give a slight generalization of Lemma 10.3. Lemma 10.3 is the particular case of Corollary 10.4 with $\eta(j,n) = \eta(j/n)$.

Corollary 10.4. *Suppose* $\eta : [0,1] \to \mathbb{R}^m$ *is a random curve,*

$$\{F(j,n) : n = 1,2,\ldots, j = 1,2,\ldots,n\}$$

are nonnegative random variables, and

$$\{\eta(j,n) : n = 1,2,\ldots, j = 1,2,\ldots,n\}$$

are $\mathbb{R}^m$*-valued random variables all defined on the same probability space. Suppose* $0 < d \leq m$*, and there exist a subpower function* ψ*,* $0 < \xi < 1$*, and* $c > 0$ *such that the following holds for* $n = 1,2,\ldots,$ *and* $1 \leq j \leq k \leq n$*:*

$$\frac{1}{c} \leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[F(j,n)] \leq c, \tag{45}$$

$$\mathbb{E}[F(j,n)F(k,n)] \leq \left(\frac{n}{k-j+1}\right)^{\xi} \psi\left(\frac{n}{k-j+1}\right), \tag{46}$$

$$|\eta(j,n) - \eta(k,n)| \geq \left(\frac{|k-j|}{n}\right)^{\frac{1-\xi}{d}} \psi\left(\frac{n}{|j-k|+1}\right)^{-1} 1\{F(j,n)F(k,n) > 0\}, \tag{47}$$

and such that with probability one

$$\lim_{n\to\infty} \max\{\operatorname{dist}[\eta(j,n), \eta[0,1]] : j = 1,\ldots,n\} = 0. \tag{48}$$

Then

$$\mathbb{P}\{\dim_h(\eta[0,1]) \geq d\} > 0.$$

In particular, if it is known that there is a d_* *such that* $\mathbb{P}\{\dim_h(\eta[0,1]) = d_*\} = 1$*, then* $d_* \geq d$*.*

Proof. The proof proceeds as in Proposition 10.3. The measure $\mu_{j,n}$ in the proof is placed on the ball centered at $\eta(j,n)$ rather than $\eta(j/n)$. The key observation is that on the event (48) any subsequential limit of the measures ν_n must be supported on $\eta[0,1]$. □

10.2. Basic idea

In order to prove (38), we will show that the conditions of Corollary 10.4 are satisfied with

$$\xi = d(d-2) + 1 = \frac{1}{16a^2} = \frac{\kappa^2}{64} \in (0,1).$$

Let

$$\beta = \frac{1}{4a} - \frac{1}{2} = \frac{\kappa}{8} - \frac{1}{2} = d - \frac{3}{2} = \frac{\xi - 1}{d} + \frac{1}{2}.$$

For fixed positive integer n and integers $1 \leq j < k \leq n$, we let

$$S = S_{j,n} = 1 + \frac{j-1}{n}, \quad T = T_{j,k,n} = \frac{k-j}{n}, \quad S+T = 1 + \frac{k-1}{n}, \tag{49}$$

$$\hat{f}_{j,n} = \hat{f}_{S_{j,n}},$$

$$\eta(t) = \gamma(1+t), \quad \eta(j,n) = \hat{f}_{j,n}(i/\sqrt{n}).$$

Note that $1 \leq S \leq S+T \leq 2, 0 \leq T \leq 1$. We will define an event $E_{j,n}$ with indicator function $I(j,n)$ and define

$$F(j,n) = n^{1-\frac{d}{2}}\, |\hat{f}'_{S_{j,n}}(i/\sqrt{n})|^d \, I(j,n).$$

The event $E_{j,n}$ will describe "typical" behavior when we weight the paths by $|\hat{f}'_{j,n}(i/\sqrt{n})|^d$; in particular, it will satisfy

$$\mathbb{E}\left[|\hat{f}'_{j,n}(i/\sqrt{n})|^d \, I(j,n)\right] \asymp \mathbb{E}\left[|\hat{f}'_{j,n}(i/\sqrt{n})|^d\right] \asymp n^{\frac{d}{2}-1}. \tag{50}$$

We define the event in Section 10.5. The typical value of $|\hat{f}'_{j,n}(i/\sqrt{n})|$ when weighted as above is n^β; more precisely, there exists a subpower function ϕ such that on the event $E_{j,n}$,

$$n^\beta\, \phi(n)^{-1} \leq |\hat{f}'_{j,n}(i/\sqrt{n})|^d \leq n^\beta\, \phi(n).$$

The "one-point" estimate (50) suffices to prove (45) in Corollary 10.4. One needs to prove the other two conditions as well. This is most easily done by considering the reverse time Loewner flow.

10.3. Reverse time

It is known (see, e.g., [10, 4]) that estimates for $\hat{f}'_t$ are often more easily derived by considering the reverse (time) Loewner flow. This is how the one-point estimate is derived. In this subsection, we review the facts about the Loewner equation in reverse time that we will need. Suppose that g_t is the solution to the Loewner equation

$$\partial_t g_t(z) = \frac{a}{g_t(z) - V_t}, \quad g_0(z) = z. \tag{51}$$

Here V_t can be any continuous function, but we will be interested in the case where V_t is a standard Brownian motion.

For fixed $T > 0$, let $F_t^{(T)}, 0 \leq t \leq T$, denote the solution to the time-reversed Loewner equation

$$\partial_t F_t^{(T)}(z) = -\frac{a}{F_t^{(T)}(z) - V_{T-t}} = \frac{a}{V_{T-t} - F_t^{(T)}(z)}, \quad F_0^{(T)}(z) = z. \tag{52}$$

Note that

$$F_{s+T}^{(S+T)}(z) = F_s^{(S)}(F_T^{(S+T)}(z)), \quad 0 \leq s \leq S.$$

Lemma 10.5. *If $t \leq T$, then $F_t^{(T)} = f_{t,T-t}$. In particular, $F_T^{(T)} = f_T$.*

Proof. Fix T, and let $u_t = F^{(T)}_{T-t}$. Then (52) implies that u_t satisfies

$$\dot{u}_t(z) = \frac{a}{u_t(z) - V_t}, \quad u_T(z) = z.$$

By comparison with (51), we can see that $u_t(z) = g_t(f_T(z))$, and one can check that $g_t \circ f_T = f_{t,T-t}$. □

We will be using the reverse-time flow, to study the behavior of $\hat{f}$ at one or two times. We leave the simple derivation of the next lemma from the previous lemma to the reader. A primary purpose of stating this lemma now is to set the notation for future sections.

Lemma 10.6. *Suppose $S, T > 0$ and $V : [0, S+T] \to \mathbb{R}$ is a continuous function. Suppose $g_t, 0 \le t \le S+T$ is the solution to* (51)*. As before, let $f_t = g_t^{-1}$ and $\hat{f}_t(z) = f_t(z + V_t)$. Let*

$$U_t = V_{S+T-t} - V_{S+T}, \quad 0 \le t \le S+T,$$
$$\tilde{U}_t = V_{S-t} - V_S = U_{T+t} - U_T, \quad 0 \le t \le S,$$

and let $h_t, 0 \le t \le S+T$, $\tilde{h}_t, 0 \le t \le S$, be the solutions to the reverse-time Loewner equations

$$\partial_t h_t(z) = \frac{a}{U_t - h_t(z)}, \quad h_0(z) = z,$$
$$\partial_t \tilde{h}_t(z) = \frac{a}{\tilde{U}_t - \tilde{h}_t(z)} = \frac{a}{U_{T+t} - U_T - \tilde{h}_t(z)}, \quad \tilde{h}_0(z) = z.$$

Then

$$\hat{f}_S(z) = \tilde{h}_S(z) - \tilde{U}_S, \quad \hat{f}_{S+T}(z) = h_{S+T}(z) - U_{S+T},$$
$$h_{S+T}(z) = \tilde{h}_S(h_T(z) - U_T) + U_T. \tag{53}$$

In particular,

$$\hat{f}'_S(w)\,\hat{f}'_{S+T}(z) = h'_T(z)\,\tilde{h}'_S(h_T(z) - U_T)\,\tilde{h}'_S(w),$$
$$\hat{f}_{S+T}(z) - \hat{f}_S(w) = \tilde{h}_S(h_T(z) - U_T) - \tilde{h}_s(w).$$

Proof. Note that

$$\partial_t[h_t(z) + V_{S+T}] = \frac{a}{V_{S+T-t} - (h_t(z) + V_{S+T})}, \quad h_0(z) + V_{S+T} = z + V_{S+T}.$$

Therefore, $f_{S+T}(z + V_{S+T}) = h_{S+T}(z) + V_{S+T} = h_{S+T}(z) - U_{S+T}$. Note also that

$$\partial_t[\tilde{h}_t(h_T(z) - U_T) + U_T] = \frac{a}{U_{T+t} - [\tilde{h}_t(h_T(z) - U_T) + U_T]},$$
$$\tilde{h}_0(h_T(z) - U_T) + U_T = h_T(z).$$

This gives (53). □

Remark. If V_t is a Brownian motion starting at the origin, then $U_t, \tilde{U}_t$ are standard Brownian motions starting at the origin. Moreover $\{U_t : 0 \le t \le T\}$ and $\{\tilde{U}_t : 0 \le t \le S\}$ are independent.

We let $\mathcal{F} = \mathcal{F}_S$ denote the σ-algebra generated by $\{V_s : s \leq S\} = \{U_{T+s} - U_T : s \leq S\}$ and $\mathcal{G} = \mathcal{G}_{S,T}$ the σ-algebra generated by $\{V_{S+t} - V_S : 0 \leq t \leq T\} = \{U_t : 0 \leq t \leq T\}$. Note that $\mathcal{F}$ and $\mathcal{G}$ are independent. We let $\mathcal{F} \vee \mathcal{G}$ be the σ-algebra generated by $\{U_t : 0 \leq t \leq S + T\}$.

Let us give an idea of the reason for including the $I(j,n)$ term in our definition of $F(j,n)$ for establishing the second bound in Corollary 10.4. Recall that

$$F(j,n) = n^{1-\frac{d}{2}} \, |\hat{f}'_{j,n}(i/\sqrt{n})|^d \, I(j,n)$$

where $I(j,n)$ is the indicator function of an event on which $|\hat{f}'_{j,n}(i/\sqrt{n})| \approx n^\beta$. To give "two-point" estimates, we need to consider $\mathbb{E}[F(j,n)F(k,n)]$. Suppose, for example, $j = k$. If we did not include the $I(j,n)$ term, then we would be estimating

$$\mathbb{E}\left[|\hat{f}'_{j,n}(i/\sqrt{n})|^{2d}\right],$$

which is not of the same order of magnitude as $(\mathbb{E}|\hat{f}'_{j,n}(i/\sqrt{n})|^d])^2$. Indeed, if we weight paths by $|\hat{f}'_{j,n}(i/\sqrt{n})|^{2d}$, we do not concentrate on paths with $|\hat{f}'_{j,n}(i/\sqrt{n})| \approx n^\beta$ but rather on paths with $|\hat{f}'_{j,n}(i/\sqrt{n})| \approx n^{\beta'}$ for some $\beta' > \beta$. However, when we include the $I(j,n)$ term, we can write roughly

$$\mathbb{E}[|\hat{f}'_S(i/\sqrt{n})|^{2d} \, I(j,n))] \approx n^{d\beta} \, \mathbb{E}[|\hat{f}'_S(i/\sqrt{n})|^d \, I(j,n)] \approx n^{2d\beta} \, \mathbb{P}(E_{j,n}).$$

10.4. Lemmas about conformal maps

In this section we collect some facts about conformal maps. They are consequences of the Koebe (1/4)-theorem and the distortion theorem (see, e.g., [5, Section 3.2]) which we now recall. Suppose $f : D \to f(D)$ is a conformal transformation, $z \in D$, and $d_{z,D} = \mathrm{dist}(z, \partial D)$. The (1/4)-theorem states that $f(D)$ contains the open ball of radius $d_{z,D} \, |f'(z)|/4$ about $f(z)$ and the distortion theorem implies

$$\frac{1-r}{(1+r)^3} \, |f'(z)| \leq |f'(w)| \leq \frac{1+r}{(1-r)^3} \, |f'(z)|, \qquad |w - z| \leq r d_{z,D}. \tag{54}$$

The immediately corollary of the (1/4)-theorem that we will need is the following: if $h : \mathbb{H} \to h(\mathbb{H})$ is a conformal transformation and $\mathrm{Im}(z) = y_z > y_w = \mathrm{Im}(w)$, then

$$|h(z) - h(w)| \geq \frac{|h'(z)| \, (y_z - y_w)}{4} \tag{55}$$

The form of the distortion theorem is the following lemma. One can values of c, α, but the actual values will not be important to us.

Lemma 10.7. *There exist $c_2, \alpha < \infty$ such that if $h : \mathbb{H} \to h(\mathbb{C})$ is a conformal transformation, $r \geq 1$, and*

$$z, w \in \mathcal{R}(r) := [-r, r] \times [1/r, r] = \{x + iy : -r \leq x \leq r, \quad 1/r \leq y \leq r\},$$

then

$$|h'(z)| \leq c_2 \, r^\alpha \, |h'(w)|.$$

Proof. The map $F(z) = (z - i)/(z + i)$ maps $\mathbb{H}$ conformally onto $\mathbb{D}$, so we can write

$$h(z) = f(F(z)),$$

where $f : \mathbb{D} \to h(D)$. We can apply the distortion theorem to f. Details are omitted. □

10.5. Defining the $F(j,n)$

In this section, we will define the event $E_{j,n}$, which will be $\mathcal{F}$-measurable, such that

$$F(j,n) = n^{1-\frac{d}{2}} \left|\hat{f}'_S(i/\sqrt{n})\right|^d 1_{E_{j,n}}. \tag{56}$$

We write

$$E_{j,n} = E_{j,n,1} \cap \cdots \cap E_{j,n,6},$$

for events that we define below. We define $F(k,n)$ similarly; it will be $(\mathcal{F} \vee \mathcal{G})$-measurable.

The event $E_{j,n}$ is defined in terms of the solution of the time-reversed Loewner equation. Let $h_t, \tilde{h}_t$ as in Lemma 10.6. We write $Z_t = h_t(i/\sqrt{n}) - U_t = X_t + iY_t, \tilde{Z}_t = \tilde{h}_t(i/\sqrt{n}) - \tilde{U}_t = \tilde{X}_t + i\tilde{Y}_t$. In particular, for $0 \le s \le S$,

$$h_{s+T}(i/\sqrt{n}) = \tilde{h}_s(Z_T) + U_T,$$

$$h'_{s+T}(i/\sqrt{n}) = \tilde{h}'_s(Z_T)\, h'_T(i/\sqrt{n}).$$

Remark. Note that the transformation h_T is $\mathcal{G}$-measurable and the transformation $\tilde{h}_S$ is $\mathcal{F}$-measurable. The random variable Z_T is $\mathcal{G}$-measurable. The random variable $\tilde{h}'_S(Z_T)$ is neither $\mathcal{F}$-measurable nor $\mathcal{G}$-measurable. The key to bounding correlations at times S and $S + T$ is handling this random variable.

The six events will depend on a subpower function ϕ_0 to be determined later. Given ϕ_0 we define the following events. Recall that $1 \le S \le S + T \le 2$,

$$E_{k,n,1} = \left\{Y_t \ge t^{\frac{1}{2}}\, \phi_0(1/t)^{-1} \text{ for } 1/n \le t \le S + T\right\}. \tag{57}$$

$$E_{k,n,2} = \left\{Y_t \ge t^{\frac{1}{2}}\, \phi_0(nt)^{-1} \text{ for } 1/n \le t \le S + T\right\}. \tag{58}$$

$$E_{k,n,3} = \left\{|X_t| \le t^{\frac{1}{2}}\, \phi_0(1/t) \text{ for } 1/n \le t \le S + T\right\},$$

$$E_{k,n,4} = \left\{|X_t| \le t^{\frac{1}{2}}\, \phi_0(nt) \text{ for } 1/n \le t \le S + T\right\}.$$

$$E_{k,n,5} = \left\{\frac{(nt)^\beta}{\phi_0(nt)} \le \left|\, h'_t(i/\sqrt{n}) \,\right| \le (nt)^\beta\, \phi_0(nt) \text{ for } 1/n \le t \le S + T\right\}$$

$$E_{k,n,6} = \left\{\frac{t^{-\beta}}{\phi_0(1/t)} \le \left|\frac{h'_{S+T}(i/\sqrt{n})}{h'_t(i/\sqrt{n})}\right| \le t^{-\beta}\, \phi_0(1/t) \text{ for } 1/n \le t \le S + T\right\}. \tag{59}$$

$E_{j,n,\cdot}$ are defined in the same way replacing $h_t, U_t, Z_t, S + T$ with $\tilde{h}_t, \tilde{U}_t, \tilde{Z}_t, S$.

Remark. What we would really like to do is to define an event of the form

$$Y_t \asymp t^{\frac{1}{2}}, \quad |X_t| \leq c\, t^{\frac{1}{2}}, \quad |h_t'(i/\sqrt{n})| \asymp (nt)^{\beta},$$

for all $0 \leq t \leq S+T$. However, this is too strong a restriction if we want the event to have positive probability (in the weighted measure). What we have done is modify this so that quantities are comparable for times near zero and for times near $S+T$ but the error may be larger for times in between (but still bounded by a subpower function).

Theorem 10.8. *There exist c_1, c_2 such that for all $t \geq 1/n$,*

$$\mathbb{E}\left[\left|h_t'(i/\sqrt{n})\right|^d\right] = \mathbb{E}\left[|h_{tn}'(i)|^d\right] \leq c_2\,(tn)^{\frac{d}{2}-1}. \tag{60}$$

Moreover there exists a power function ϕ_0 such that if $E_{j,n}$ is defined as above, then

$$\mathbb{E}\left[\left|\tilde{h}_S'(i/\sqrt{n})\right|^d I(j,n)\right] \geq c_1\; n^{\frac{d}{2}-1}. \tag{61}$$

Proof. See Theorems 8.1 and 9.1. □

Remark. The equality in (60) follows immediately from scaling. Since $|\tilde{h}_S'(i/\sqrt{n})|^d \approx n^{\beta}$ on $E_{j,n}$, we can see that

$$n^{\frac{d}{2}-1-d\beta}\,\phi(n)^{-1} \leq \mathbb{P}(E_{j,n}) \leq n^{\frac{d}{2}-1-d\beta}\,\phi(n)^{-1}.$$

Before proceeding further, we note that the Loewner equation gives

$$dX_t = -\frac{a\,X_t}{|Z_t|^2}\,dt - dU_t,$$

which implies

$$|U_t + X_t| \leq \int_0^t \frac{a\,|X_s|\,ds}{|Z_s|^2}.$$

Using this we can show that on the event $E_{k,n}$,

$$|U_t| \leq t^{\frac{1}{2}}\,\phi_0(1/t) \text{ for } 1/n \leq t \leq S+T, \tag{62}$$

$$|U_t| \leq t^{\frac{1}{2}}\,\phi_0(nt) \text{ for } 1/n \leq t \leq S+T,$$

with perhaps a different choice of subpower function ϕ_0. Hence, we may assume that the function ϕ_0 is chosen so that the last two inequalities hold as well.

10.6. Handling the correlations

Theorem 10.8 discusses the function $\tilde{h}_t$ and the corresponding processes $\tilde{X}_t, \tilde{Y}_t$ for a fixed value of S. In this section we assume Theorem 10.8 and show how to verify the hypotheses of Corollary 10.4 for ξ as defined earlier and some subpower function ϕ. Here $F(j,n)$ is defined as in (56). The first hypothesis (45) follows immediately from (60) so we will only need to consider (46)–(48). Throughout this subsection ϕ will denote a subpower function, but its value may change from line to line.

10.6.1. The estimate (46)**.** We first consider $j = k$. Then

$$\mathbb{E}[F(j,n)^2] = n^{2-d}\,\mathbb{E}\left[\left|\tilde{h}_S'(i/\sqrt{n})\right|^{2d}\, I(j,n)\right].$$

On the event $E_{j,n}$ we know that $|\tilde{h}_S'(i/\sqrt{n})| \le n^{\beta}\,\phi(n)$. Therefore, using (60),

$$\mathbb{E}[F(j,n)^2] \le n^{2-d+\beta d}\,\mathbb{E}\left[\left|\tilde{h}_S'(i/\sqrt{n})\right|^{d}\right]\,\phi(n) \le n^{\xi}\,\phi(n).$$

We now assume $j < k$. We need to give an upper bound for

$$\mathbb{E}[F(j,n)\,F(k,n)] = n^{2-d}\,\mathbb{E}\left[\,1_{E_{j,n}}\,|\tilde{h}_S'(i/\sqrt{n})|^d\,1_{E_{k,n}}\,|h_{S+T}'(i/\sqrt{n})|^d\,\right].$$

Let $\tilde{E}_{k,n} = \tilde{E}_{k,n,1} \cap \tilde{E}_{k,n,3}$ where $\tilde{E}_{k,n,j}$ is defined as $E_{k,n,j}$ except that $1/n \le t \le S+T$ is replaced with $1/n \le t \le T$. Then $\tilde{E}_{k,n}$ is $\mathcal{G}$-measurable and $E_{k,n} \subset \tilde{E}_{k,n}$. Using (53), we can write

$$h_{T+S}'(i/\sqrt{n}) = h_T'(i/\sqrt{n})\,\tilde{h}_S'(Z_T).$$

Therefore,

$$n^{d-2}\,\mathbb{E}[F(j,n)\,F(k,n)] \le$$
$$\mathbb{E}\left[\,1_{E_{j,n}}\,|\tilde{h}_S'(i/\sqrt{n})|^d\,|\tilde{h}_S'(Z_T)|^d\,1_{\tilde{E}_{k,n}}\,|h_T'(i/\sqrt{n})|^d\,\right]. \qquad (63)$$

This is the expectation of a product of five random variables. The first two are $\mathcal{F}$-measurable and the last two are $\mathcal{G}$-measurable. The middle random variable $|\tilde{h}_S'(Z_T)|$ uses information from both σ-algebras: the transformation $\tilde{h}_S$ is $\mathcal{F}$-measurable but it is evaluated at Z_T which is $\mathcal{G}$-measurable.

We claim that it suffices to show that on the event $E_{j,n} \cap \tilde{E}_{k,n}$,

$$\left|\tilde{h}_S'(Z_T)\right|^d \le T^{-\beta d}\,\phi(1/T), \qquad (64)$$

for some subpower function ϕ. Indeed, once we have established this we can see that the expectation in (63) is bounded above by

$$T^{-\beta d}\,\phi(1/T)\,\mathbb{E}\left[|\tilde{h}_S'(i/\sqrt{n})|^d\,\left|h_T'(i/\sqrt{n})\right|^d\,\right],$$

which by independence equals

$$T^{-\beta d}\,\phi(1/T)\,\mathbb{E}\left[|\tilde{h}_S'(i/\sqrt{n})|^d\right]\,\mathbb{E}\left[\,|h_T'(i/\sqrt{n})|^d\,\right].$$

Using (60), we then have that this is bounded by

$$T^{-\beta d}\,\phi(T)\,n^{\frac{d}{2}-1}\,(nT)^{\frac{d}{2}-1} = T^{-\xi}\,\phi(1/T) = \left(\frac{n}{k-j}\right)^{\xi}\,\phi\left(\frac{n}{k-j}\right).$$

Hence, we only need to establish (64).

Let $w = Z_T$. By the definition of $\tilde{E}_{k,n}$, there is a subpower function ϕ such that $w \in T^{1/2}\,\mathcal{R}(\phi(1/T))$, where $\mathcal{R}(r)$ is as defined in Lemma 10.7. Using the

Loewner equation and the fact that the imaginary part increases in the reverse flow, we can see that

$$\tilde{h}_T(w) \in T^{1/2}\,\mathcal{R}(\phi(1/T)), \quad |\tilde{h}'_T(w)| \leq \phi(1/T).$$

for perhaps a different ϕ (we will change the value of ϕ from line to line). By the definition of $E_{j,n}$ and (62), we know that

$$\tilde{Z}_T \in T^{1/2}\,\mathcal{R}(\phi(1/T)), \qquad |\tilde{U}_T| \leq T^{1/2}\,\phi(1/T).$$

Hence also,

$$\tilde{h}_T(w) - \tilde{U}_T \in T^{1/2}\,\mathcal{R}(\phi(1/T)).$$

If we define $\tilde{h}_{T,S}$ by $\tilde{h}_S(z_1) = \tilde{h}_{T,S}(\tilde{h}_T(z_1) - \tilde{U}_T)$, then we know by the definition of $E_{j,n}$ that

$$|\tilde{h}'_{T,S}(\tilde{Z}_S)| \leq T^{-\beta}\,\phi(1/T).$$

Hence, by Lemma 10.7,

$$|\tilde{h}_{T,S}(\tilde{h}_T(w) - \tilde{U}_T)| \leq T^{-\beta}\,\phi(1/T).$$

10.6.2. The estimate (47)**.** Assume that $j < k$ and that $E_{j,n}$ and $E_{k,n}$ both occur. Then,

$$\eta(k,n) = \hat{f}_{S+T}(i/\sqrt{n}) = h_{S+T}(1/\sqrt{n}) + U_{T+S} = \tilde{h}_S(Z_T) + U_S.$$

Therefore,

$$\eta(k,n) - \eta(j,n) = \tilde{h}_S(Z_T) - \tilde{h}_S(i/\sqrt{n}).$$

Using the Loewner equation, we see that there is a c_1 such that

$$Y_T - n^{-1/2} > c_1\,Y_T. \tag{65}$$

(Since Y_T is increasing we only need to check this for $T = 1/n$.) Using (55), we get

$$|\eta(k,n) - \eta(j,n)| = \left|\tilde{h}_S(Z_T) - \tilde{h}_S(1/\sqrt{n})\right| \geq \frac{c_1}{4}\,Y_T\,|\tilde{h}'_S(Z_T)|.$$

Hence on our event,

$$|\eta\,(k,n) - \eta\,(j,n)|^d \geq T^{\frac{d}{2}+d\beta}\,\phi\left(\frac{n}{k-j}\right)^{-1} = T^{\xi-1}\,\phi\left(\frac{n}{k-j}\right)^{-1}.$$

Remark. Note that we do not expect that last estimate to hold for all k, j, especially for $\kappa > 4$ for which SLE_κ has double points. The restriction to the event $E_{j,n} \cap E_{k,n}$ is a major restriction.

10.6.3. The estimate (48)**.** This estimate was essentially proved by Rohde and Schramm [10] when they proved existence of the curve. In fact, we can give an argument here. On the event $E_{j,n}$, we have $|\hat{f}'_S(i/\sqrt{n})| \approx n^\beta$. Therefore, using the Koebe (1/4)-Theorem we can conclude on this event that for every $\epsilon > 0$, there is a c such that

$$\text{dist}\left[\hat{f}_S(i/\sqrt{n}), \gamma[0,s] \cap \mathbb{R}\right] \leq c\,n^{\beta+\epsilon}\,n^{-1/2} = c\,n^{d+\epsilon-2}.$$

Since $d < 2$, this goes to 0 for ϵ sufficiently small.

Appendix A. On the Girsanov theorem

In this section, we give a review of the Girsanov theorem. Assume B_t is a standard Brownian motion and $\mathcal{F}_t = \sigma\{B_s : 0 \leq s \leq t\}$ is the filtration generated by the Brownian motion. Suppose that M_t is a nonnegative solution of the stochastic differential equation

$$dM_t = M_t\, A_t\, dB_t, \quad M_0 = x_0 > 0. \tag{66}$$

Here, and throughout, A_t will denote a process such that A_t is $\mathcal{F}_t$-measurable and of the form

$$A_t = \tilde{A}_t\, 1\{T > t\},$$

for some stopping time T and continuous process $\tilde{A}_t$. More precisely, we require (66) to hold only for $M_t > 0$. and if $\tau = \inf\{s : M_s = 0\}$, then $M_t = M_\tau \equiv 0$ for $t \geq \tau$.

Solutions to (66) are not necessarily martingales because it is possible for mass to "escape to infinity in finite time". However solutions to (66) are nonnegative continuous *supermartingales*,

$$\mathbb{E}[M_t \mid \mathcal{F}_s] \leq M_s, \quad s \leq t.$$

In order to show that M_t is a martingale it suffices to show that for each t, $\mathbb{E}[M_t] = x_0$, for then for each $s < t$

$$\mathbb{E}\left[M_t - \mathbb{E}[M_t \mid \mathcal{F}_s]\right] = 0.$$

Since the integrand is nonnegative this implies that $M_t = \mathbb{E}[M_t \mid \mathcal{F}_s]$.

A sufficient condition for a continuous local martingale to be a martingale is boundedness in the following sense: for each t there exists $K_t < \infty$ such that with probability one $|M_s| \leq K_t, 0 \leq s \leq t$.

If M_t is a nonnegative continuous local submartingale, we can obtain a martingale by stopping the process. To be more precise, let $T_n = \inf\{t : M_t \geq n\}$. Then $M_{t\wedge T_n}$ is a martingale. Suppose we know that for each t,

$$\lim_{n\to\infty} \mathbb{E}\left[M_{t\wedge T_n}\,;\, T_n < t\right] = 0. \tag{67}$$

Then the optional sampling theorem and the monotone convergence theorem imply

$$\mathbb{E}[M_t] = \mathbb{E}[M_0].$$

The condition (67) means that mass does not escape to infinity by time t. Therefore, to show that a nonnegative continuous local martingale is a martingale it suffices to prove (67) for each t.

The Girsanov theorem is a theorem about nonnegative *martingales* satisfying (66). Suppose M_t is such a martingale. Let $\mathbf{Q}_t$ be the probability measure on $\mathcal{F}_t$-measurable events given by

$$\mathbf{Q}_t(V) = M_0^{-1}\, \mathbb{E}\left[M_t\, 1_V\right].$$

If $s < t$ and V is $\mathcal{F}_s$-measurable, then it is also $\mathcal{F}_t$-measurable and $\mathbf{Q}_s(V) = \mathbf{Q}_t(V)$. This follows from properties of conditional expectation:

$$\mathbb{E}\left[M_t\, 1_V\right] = \mathbb{E}\left[\mathbb{E}(M_t\, 1_V \mid \mathcal{F}_s)\right] = \mathbb{E}\left[1_V\, \mathbb{E}(M_t \mid \mathcal{F}_s)\right] = \mathbb{E}\left[M_s\, 1_V\right].$$

(Note that this calculation uses the fact that M_t is a martingale.) Hence these measures are consistent and give rise to a measure on $\mathcal{F}_\infty$ which we call $\mathbf{Q}$.

Theorem A.1 (Girsanov). *Under the assumptions above, with respect to* $\mathbf{Q}$,

$$W_t = B_t - \int_0^t A_s\, ds \tag{68}$$

is a standard Brownian motion.

We can write (68) as

$$dB_t = A_t\, dt + dW_t. \tag{69}$$

We can state the theorem informally as: "*if we weight the paths by the martingale M_t, then in the weighted measure B_t satisfies* (69)".

The Girsanov theorem requires that M_t be a martingale. For many applications in SLE, one only knows that M_t satisfies (66), and hence only that M_t is a *local* martingale. However, we can still use Girsanov by using stopping times (this procedure is sometimes called localization). Assume for ease that $M_0 = 1$; otherwise, we consider M_t/x_0. Let $T = T_n = \inf\{t : M_t \geq n \text{ or } |A_t| \geq n\}$. Then $M_{t\wedge T}$ is a bounded martingale satisfying (66) which we write as

$$dM_{t\wedge T} = M_{t\wedge T}\, A_t\, dB_t, \quad 0 \leq t \leq T.$$

The Girsanov theorem applies to $M_{t\wedge T}$. Hence, if we weight by $M_{t\wedge T}$, then in the weighted measure B_t satisfies

$$dB_t = A_t\, dt + dW_t, \quad 0 \leq t \leq T. \tag{70}$$

As a slight abuse of notation, we can say that if we weight by the local martingale M_t, then in the weighted measure B_t satisfies (69). This is shorthand for saying that (70) holds for all $n < \infty$. To show that M_t is a martingale, it suffices to establish (67), and do this one can use (70). If we let $\mathbf{Q}$ denote the weighted measure, we can rewrite (67) as

$$\lim_{n\to\infty} \mathbf{Q}\{T_n < t\} = 0.$$

Remark. The Girsanov theorem could be stated as a theorem about nonnegative local martingales in which case the measure $\mathbf{Q}_t$ on $\mathcal{F}_t$ is a subprobability measure. We find it more convenient to restrict the Girsanov theorem to martingales and to use stopping times.

As an example, suppose that B_t is a standard Brownian motion and $\phi(z) = e^{\psi(z)}$ is a positive C^2 function. Then Itô's formula shows that

$$M_t = \phi(B_t)\, \exp\left\{-\frac{1}{2}\int_0^t [\psi''(B_s) + \psi'(B_s)^2]\, ds\right\},$$

is a local martingale satisfying

$$dM_t = \psi'(B_t)\, M_t\, dB_t.$$

If we weight by the local martingale M_t, then

$$dB_t = \psi'(B_t)\, dt + dW_t, \tag{71}$$

where W_t is a Brownian motion in the new measure which we denote by $\mathbf{Q}$. This must be interpreted in terms of stopping times; however, if solutions to the SDE (71) do not blow up in finite time, then M_t does not blow up in finite time in the measure $\mathbf{Q}$ and hence M_t is a martingale. A sufficient condition for M_t to be a martingale is that ψ' is uniformly bounded.

Assume M_t is a martingale. Let $p_t(x, y)$ denote the transition density for the Brownian motion and let $q_t(x, y)$ denote the transition density for the SDE (71). We claim that for all t, x, y,

$$\phi(x)^2\, q_t(x, y) = \phi(y)^2\, q_t(y, x). \tag{72}$$

To see this from the perspective of the Girsnaov theorem, consider the set of paths $\gamma : [0, t] \to \mathbb{R}$ with $\gamma(0) = x, \gamma(t) = y$. Then $\mathbb{P}, \mathbf{Q}$ give two measures on the set of such paths with

$$\frac{d\mathbf{Q}}{d\mathbb{P}}\,(\gamma) = \frac{\phi(y)}{\phi(x)} \exp\left\{-\frac{1}{2}\int_0^t [\psi'' + (\psi')^2](\gamma(s))\, ds\right\}.$$

Here we have used the fact that M_t is a deterministic function of the path $\gamma(s), 0 \leq s \leq t$. Similarly, if we consider the reversed path γ^R with $\gamma^R(s) = \gamma(t - s)$, we get

$$\frac{d\mathbf{Q}}{d\mathbb{P}}\,(\gamma^R) = \frac{\phi(x)}{\phi(y)} \exp\left\{-\frac{1}{2}\int_0^t [\psi'' + (\psi')^2](\gamma^R(s))\, ds\right\}.$$

The integral inside the exponential is not easy to compute but it is the same for γ and γ^R, which gives

$$\phi(x)^2\, \frac{d\mathbf{Q}}{d\mathbb{P}}\,(\gamma) = \phi(y)^2\, \frac{d\mathbf{Q}}{d\mathbb{P}}\,(\gamma^R).$$

Combining this with the symmetry of $\mathbb{P}$ gives (72). In particular,

$$\int \phi^2(x)\, q_t(x, y)\, dx = \int \phi^2(y)\, q_t(y, x)\, dx = \phi^2(y).$$

This shows that ϕ^2 us an invariant density for (72).

Often one starts with the SDE (71) in which case one can find ϕ using

$$\phi = e^{\int \psi'}.$$

Example. If $\psi(x) = bx$, we obtain the well-known exponential martingale

$$M_t = e^{bB_t}\, e^{-b^2 t/2}.$$

In the weighted measure, which we denote by $\mathbf{Q}_b$, B_t satisfies

$$dB_t = b\, dt + dW_t,$$

i.e., B_t is a Brownian motion with drift. This has no explosion in finite time, so M_t is a martingale and

$$e^{-b^2t/2} = \mathbb{E}[M_t\, e^{-bB_t}] = \mathbb{E}_{\mathbf{Q}_b}[e^{-bB_t}] = e^{-b^2t}\, \mathbb{E}_{\mathbf{Q}_b}[e^{-bW_t}].$$

This is the standard martingale computation for the moment generating function for a normal.

Example. Assume $B_0 > 0$ and let $\phi(x) = x^b$, i.e., $\psi(x) = b\,\log x$. Then

$$M_t = B_t^b \, \exp\left\{ -\frac{b(b-1)}{2} \int_0^t \frac{ds}{B_s^2} \right\}.$$

In the weighted measure, B_t satisfies the Bessel equation

$$dB_t = \frac{b}{B_t}\, dt + dW_t.$$

If $b \geq 1/2$, then the process under the weighted measure does not reach 0 in finite time. For these values, we can say that M_t is a martingale.

Example. If $\psi(x) = -bx^2/2$, then

$$M_t = e^{-bB_t^2/2}\, e^{-bt}\, \exp\left\{ b^2 \int_0^t B_s^2\, ds \right\}.$$

In the weighted measure, B_t satisfies the Ornstein-Uhlenbeck equation

$$dB_t = -b\, B_t\, dt + dW_t.$$

The invariant density is proportional to $\phi(x)^2 = e^{-bx^2}$ from which we see that the process is positive recurrent.

Example. In the study of radial SLE, the case $\psi'(x) = (b/2)\cot(x/2)$ arises for which

$$\phi(x) = \exp\left\{ \int \psi' \right\} = \sin^b(x/2).$$

Assume $B_0 \in (0, 2\pi)$. In the weighted measure B_t satisfies

$$dB_t = \frac{b}{2}\, \cot(B_t/2)\, dt + dW_t.$$

By comparison with a Bessel process we see that this process stays in $(0, 2\pi)$ for all time provided that $b \geq 1/2$. The invariant density is $\sin^{2b}(x/2)$.

Example. In Section 7, we consider the case

$$\phi(x) = [\cosh x]^b, \qquad \psi'(x) = b\, \tanh x.$$

Note that ψ' is bounded. The invariant density is $\phi(x)^2 = [\cosh x]^{2b}$.

References

[1] V. Beffara (2008). The dimension of the SLE curves, Annals of Probab. **36**, 1421–1452.

[2] D. Beliaev and S. Smirnov, Harmonic measure and SLE, preprint.

[3] K. Falconer (1990). *Fractal Geometry: Mathematical Foundations and Applications.* John Wiley & Sons.

[4] N.-G. Kang (2007). Boundary behavior of SLE, Journal of AMS, **20**, 185–210.

[5] G. Lawler (2005). *Conformally Invariant Processes in the Plane*, Amer. Math. Soc.

[6] G. Lawler, Schramm-Loewner evolution, notes for course at 2007 Park City – Institute for Advanced Study workshop, to appear.

[7] G. Lawler and S. Sheffield, The natural parametrization for the Schramm-Loewner evolution, in preparation.

[8] G. Lawler and F. Johansson, Tip multifractal spectrum for the Schramm-Loewner evolution, in preparation

[9] J. Lind (2008). Hölder regularity of the SLE trace, Trans. AMS **360**, 3557–3578.

[10] S. Rohde and O. Schramm (2005). Basic properties of SLE, Annals of Math. **161**, 879–920.

[11] O. Schramm (2000). Scaling limits of loop-erased random walks and uniform spanning trees, Israel J. Math. **118**, 221–288.

Gregory F. Lawler
Department of Mathematics
University of Chicago
5734 S. University Ave.
Chicago, IL 60637-1546, USA
e-mail: `lawler@math.uchicago.edu`

Part 3

Random Fractal Processes

Progress in Probability, Vol. 61, 111–141
© 2009 Birkhäuser Verlag Basel/Switzerland

From Fractals and Probability to Lévy Processes and Stochastic PDEs

Davar Khoshnevisan

In memory of Oded Schramm (1961–2008)

Abstract. We present a few choice examples that showcase how the topics in the title are deeply interconnected. Although this is mainly a survey article, it is almost entirely self-contained with complete, albeit somewhat stylized, proofs.

Mathematics Subject Classification (2000). Primary 28A78, 28A80, 60J30, 60H15; Secondary 35R60.

Keywords. Fractals, Lévy processes, stochastic partial differential equations.

1. Introduction

It has been known for a long time that the theories of fractals and probability are related. The principle aim of this paper is to make a case for the assertion that those theories are in fact inextricably intertwined. And that, frequently, one learns a good deal by adopting this particular point of view.

I aim to make this case by presenting a series of examples via a self-contained flow of ideas that starts with a classical problem about the ternary Cantor set, and progresses to more modern examples from stochastic analysis that are rooted in statistical mechanics. The approach presented here will have some novelty, even for the classical examples.

The only strict prerequisites to reading this article are a modest knowledge of measure-theoretic probability and a little harmonic analysis on compact abelian groups [for §3.5]. But it might help to know also some stochastic-process theory, "hard" harmonic analysis, and fractal analysis.

And now we begin at the beginning, and without further ado.

Research supported in part by a grant from the National Science Foundation grant DMS-0706728.

1.1. The Minkowski dimension

We begin our discussion by reviewing some basic facts from fractals and geometric measure theory; more detail can be found in the excellent books of Falconer [12] and Mattila [35].

One of the simplest notions of dimension is the [upper] *Minkowski dimension*, or *box dimension*, of a bounded set $G \subset \mathbf{R}^d$.

Recall that a cube $U \subset \mathbf{R}^d$ is *dyadic of side* 2^{-n} if it has the form

$$U := \left(\frac{j_1}{2^n}, \frac{j_1+1}{2^n} \right] \times \cdots \times \left(\frac{j_d}{2^n}, \frac{j_d+1}{2^n} \right], \tag{1.1}$$

for some $j := (j_1, \ldots, j_d) \in \mathbf{Z}^d$. The collection of such cubes is denoted by $\mathcal{D}_n$, and $\mathcal{D} := \cup_{n=-\infty}^{\infty} \mathcal{D}_n$ denotes the collection of all *dyadic cubes* in $\mathbf{R}^d$. On a very few occasions I might refer to *b-adic cubes of side* b^{-n}, where $b \geq 2$ is an integer; those are defined also as above, but 2^n is replaced everywhere by b^n.

The *Minkowski dimension* of a bounded set $G \subset \mathbf{R}^d$ is defined as

$$\overline{\dim}_{\mathrm{M}} G := \limsup_{n\to\infty} \frac{1}{n} \log_2 \left(\# \left\{ U \in \mathcal{D}_n : \, G \cap U \neq \varnothing \right\} \right), \tag{1.2}$$

where "$\log_2$" denotes the base-two logarithm.[1] It is possible to see that $\overline{\dim}_{\mathrm{M}}$ does not depend on the base. That is, if we use b-adic cubes in place of dyadic ones and apply $\log_b$ in place of $\log_2$, then we obtain the same numerical value for $\overline{\dim}_{\mathrm{M}} G$.

Minkowski dimension is also known as the *upper Minkowski* [or box] *dimension.* The corresponding *lower Minkowski* [or box] *dimension* is defined as

$$\underline{\dim}_{\mathrm{M}} G := \liminf_{n\to\infty} \frac{1}{n} \log_2 \left(\# \left\{ U \in \mathcal{D}_n : \, G \cap U \neq \varnothing \right\} \right). \tag{1.3}$$

Clearly $\underline{\dim}_{\mathrm{M}} G \leq \overline{\dim}_{\mathrm{M}} G$.

Minkowski dimension is easy to use [and frequently easy to compute], but has the drawback that there are countable sets of positive Minkowski dimension. An example is $G := \{1, 1/2, 1/3, \ldots\}$, whose Minkowski dimension is 1/2. Viewed from this perspective, Hausdorff dimension is a more attractive notion of dimension. We describe that notion next.

1.2. Net measures and Hausdorff dimension

Given a number $q \in (0, \infty)$ and a set $G \subset \mathbf{R}^d$, we can define the quantity $\mathcal{N}_q^n(G)$ as $\inf \sum_{k=1}^{\infty} (\text{side } I_k)^q$, where the infimum is taken over all dyadic cubes $I_1, I_2, \ldots$ that have side length $\leq 2^{-n}$ and cover G in the sense that $G \subseteq \cup_{k=1}^{\infty} I_k$. The *q-dimensional net measure of* G is the monotonic limit,

$$\mathcal{N}_q(G) := \lim_{n\to\infty} \mathcal{N}_q^n(G). \tag{1.4}$$

[1] As far as I know, Minkowski himself did not study this notion of dimension, but he did introduce the closely-related Minkowski content in geometry.

The restriction of $\mathcal{N}_q$ to the Borel sets of $\mathbf{R}^d$ – still denoted by $\mathcal{N}_q$ – is a bona fide Borel measure. The *Hausdorff dimension* $\dim_{\mathrm{H}} G$ of $G \subset \mathbf{R}^d$ is then defined unambiguously as

$$\dim_{\mathrm{H}} G := \sup\{q > 0 : \mathcal{N}_q(G) > 0\} = \inf\{q > 0 : \mathcal{N}_q(G) < \infty\}. \tag{1.5}$$

It can be shown that Hausdorff dimension has the following regularity property: For all Borel sets $G_1, G_2, \ldots \subset \mathbf{R}^d$,

$$\dim_{\mathrm{H}} \left(\bigcup_{n=1}^{\infty} G_n \right) = \sup_{n \geq 1} \dim_{\mathrm{H}} G_n. \tag{1.6}$$

In particular, $\dim_{\mathrm{H}} G = 0$ whenever G is countable.

If we use b-adic cubes in place of dyadic ones, then we obtain net measures that are, to within constant multiples, the same as $\mathcal{N}_q$. Thus, Hausdorff dimension does not depend on the base b that is used.

It can be verified that

$$\dim_{\mathrm{H}} G \leq \underline{\dim}_{\mathrm{M}} G \leq \overline{\dim}_{\mathrm{M}} G. \tag{1.7}$$

Either, or both, of these inequalities can be strict.

1.3. Riesz capacity

Typically one computes $\dim_{\mathrm{H}} G$ by separately deriving an upper and a lower bound for it. It is not hard to derive an upper bound in many cases: One strives to construct a "good dyadic cover" $\{U_j^n\}_{j=1}^{\infty}$ of side length $\leq 2^{-n}$, and then estimate $\mathcal{N}_q^n(G)$ from the above by $\sum_{k=1}^{\infty}(\text{side } U_k^n)^q$. And then one tries to find a value of q that ensures that the said sum is bounded uniformly in n; thus, $\dim_{\mathrm{H}} G \leq q$ for such a value of q.

It is more difficult to obtain lower bounds, since we have to consider all possible covers. As it turns out, there is a potential-theoretic method that is particularly well suited to obtaining lower bounds for $\dim_{\mathrm{H}} G$ in many cases.

If μ is a Borel measure on $\mathbf{R}^d$ and q is a positive constant, then the *q-dimensional Riesz energy* of μ is defined as

$$I_q(\mu) := \iint \frac{\mu(\mathrm{d}x)\, \mu(\mathrm{d}y)}{\|x - y\|^q}, \tag{1.8}$$

where $\|\cdots\|$ denotes the usual Euclidean norm. The quantity $I_q(\mu)$ might, or might not, be finite. The following result shows us how we can try to find a lower bound for $\dim_{\mathrm{H}} G$. Here and throughout, $\mathcal{P}(\cdots)$ denotes the collection of all Borel probability measures that have compact support and satisfy $\mu(G) = 1$.

Theorem 1.1 (Frostman [15]**).** *If there exist $\mu \in \mathcal{P}(G)$ and $q > 0$ such that $I_q(\mu) < \infty$, then $\dim_{\mathrm{H}} G \geq q$. Conversely, if $I_q(\mu) = \infty$ for all $\mu \in \mathcal{P}(G)$, then $\dim_{\mathrm{H}} G \leq q$.*

Define the *q-dimensional capacity* of a Borel set $G \subset \mathbf{R}^d$ as

$$\mathrm{Cap}_q(G) := \frac{1}{\inf_{\mu \in \mathcal{P}(G)} I_q(\mu)}, \tag{1.9}$$

where $\inf\varnothing := \infty$ and $1/\infty := 0$. Frostman's theorem implies that

$$\dim_{\mathrm{H}} G = \sup\left\{q>0:\ \mathrm{Cap}_q(G)>0\right\} = \inf\left\{q>0:\ \mathrm{Cap}_q(G)=0\right\}. \tag{1.10}$$

In particular, "capacity dimension = Hausdorff dimension." I demonstrate the easier – and more useful – half of Theorem 1.1 next.

Half of the proof. We suppose that $I_q(\mu)<\infty$ for some $q>0$ and $\mu\in\mathcal{P}(G)$, and prove that $\dim_{\mathrm{H}} G\geq q$, as a consequence.

For all $\epsilon>0$ we can find a dyadic cover $\{U_j\}_{j=1}^\infty$ of G such that $\sum_{j=1}^\infty |\text{side } U_j|^q$ is at most $\mathcal{N}_q(G)+\epsilon$. We fix this cover in mind, and use it as follows:

$$I_q(\mu) \geq \sum_{j=1}^\infty \iint_{U_j\times U_j} \frac{\mu(\mathrm{d}x)\,\mu(\mathrm{d}y)}{\|x-y\|^q} \geq \text{const}\cdot\sum_{j=1}^\infty \frac{[\mu(U_j)]^2}{|\text{side } U_j|^q}. \tag{1.11}$$

Since $\{U_j\}_{j=1}^\infty$ is a cover for G and μ is supported on G, Jensen's inequality implies that for all positive numbers $a_1,a_2,\ldots,$

$$\sum_{j=1}^\infty \frac{1}{a_j}\,\mu(U_j) \geq \left[\sum_{j=1}^\infty a_j\,\mu(U_j)\right]^{-1}. \tag{1.12}$$

We can set $a_j := |\text{side } U_j|^q/\mu(U_j)$ to find that

$$I_q(\mu) \geq \frac{\text{const}}{\sum_{j=1}^\infty |\text{side } U_j|^q} \geq \frac{\text{const}}{\mathcal{N}_q(G)+\epsilon}. \tag{1.13}$$

We let $\epsilon\to 0$ and then optimize over all $\mu\in\mathcal{P}(G)$ such that $I_q(\mu)$ is finite to find that $\mathcal{N}_q(G)\geq \text{const}\cdot\mathrm{Cap}_q(G)>0$. Consequently, positive capacity implies positive net measure, whence the result. □

2. Some instructive examples

2.1. The ternary Cantor set

Every $x\in[0\,,1]$ can be written as $x=\sum_{j=1}^\infty x_j 3^{-j}$ where the digits x_j are 0, 1, or 2. The ternary Cantor set C can be viewed as the collection of all points in $[0\,,1]$ whose ternary digits are in $\{0\,,2\}$. Ours differs from the more popular definition of C by at most a countable collection of points; consequently the two definitions lead to the same Hausdorff dimension. Its numerical value is contained within the following famous result of Hausdorff.

Theorem 2.1 (Hausdorff [19]**).** $\dim_{\mathrm{H}} \mathrm{C} = \log_3 2$.

Proof. The upper bound is derived by a standard covering argument, which I omit. Next you will find a proof of the lower bound that highlights some of Theorem 2.1's deep connections to probability theory: Let $\{X_j\}_{j=1}^\infty$ denote a collection of independent random variables, each taking the values 0 or 2 with probability 1/2. Then the jth ternary digit of $X := \sum_{j=1}^\infty X_j 3^{-j}$ is X_j. Clearly, $\mu(A) := \mathrm{P}\{X\in A\}$

defines a Borel probability measure on C.[2] It suffices to prove that $I_q(\mu) < \infty$ for all $q < \log_3 2$.

We might observe that if Y is independent of X and has the same distribution μ as X, then $I_q(\mu) = \mathrm{E}(|X - Y|^{-q})$, where E denotes expectation. Let Y_j denote the jth ternary digit of Y, and consider

$$J := \inf\left\{j \geq 1 : X_j \neq Y_j\right\}. \tag{2.1}$$

The triangle inequality shows that $|X-Y| \geq 3^{-J}$, whence $I_q(\mu) \leq \mathrm{E}(3^{qJ})$. Because $\mathrm{P}\{J = j\} = 2^{-j}$ for all integers $j \geq 1$, $\mathrm{E}(3^{qJ}) = \sum_{j=1}^{\infty} 3^{qj}2^{-j}$ is finite if and only if $q < \log_3 2$, and the theorem follows. □

2.2. Non-normal numbers

We follow Borel [5] and say that a number $x \in (0\,,1]$ is *simply normal in base* 2 if

$$f(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{0\}}(x_j) = \frac{1}{2}, \tag{2.2}$$

where x_j denotes the jth digit in the binary expansion of x.[3]

The celebrated *normal-number theorem* of Borel [5] asserts that Lebesgue-almost all $x \in (0\,,1]$ are simply normal in base 2 [and a little more, in fact]. Next is Borel's ingenious proof; it is simplified thanks to more than a century of afterthought.

Let $X := \sum_{j=1}^{\infty} X_j 2^{-j}$, where $\{X_j\}_{j=1}^{\infty}$ are independent random variables, each distributed uniformly on $\{0\,,1\}$. Then one verifies easily that X is distributed uniformly on $[0\,,1]$; that is, $\mathrm{P}\{X \in A\}$ is the Lebesgue measure of $A \subseteq [0\,,1]$. By the strong law of large numbers, $\mathrm{P}\{X \in \mathrm{N}_{1/2}\} = 1$ where $\mathrm{N}_{1/2}$ denotes the collection of all $x \in (0\,,1]$ that are simply normal in base 2. We have shown that $\mathrm{N}_{1/2}$ has full measure, and this completes the proof!

It turns out that many interesting nonnormal numbers form fractal collections. For instance, choose and fix a number $p \in (0\,,1)$ and define

$$\mathrm{N}_p := \left\{x \in (0\,,1] : f(x) = p\right\}. \tag{2.3}$$

Thus, the elements of N_p are numbers whose digits have the prescribed asymptotic frequencies p and $1-p$. The following striking example was conjectured originally by I.J. Good and resolved ultimately by H.G. Eggleston [11]. This example highlights some of the connections between fractals, probability theory, and notions from statistical mechanics.

Theorem 2.2 (Eggleston [11]**).** $\dim_{\mathrm{H}} \mathrm{N}_p = H(p)$, *where*

$$H(p) := p \log_2\left(\frac{1}{p}\right) + (1-p)\log_2\left(\frac{1}{1-p}\right). \tag{2.4}$$

[2] In fact, $\mu(\bullet)$ is the restriction of the net measure $\mathcal{N}_{\log_3(2)}(\bullet)$ to C.

[3] The x_j's are clearly defined uniquely for all but a countable collection of $x \in [0\,,1]$; as that collection has zero Lebesgue measure, we can safely not worry about it.

Proof. Let $X_1, X_2, \ldots$ be independent random variables, each taking the values zero and one with respective probabilities p and $1-p$. Then, $X := \sum_{j=1}^{\infty} X_j 2^{-j}$ is a random number in $(0\,,1]$ that satisfies $\mathrm{P}\{X \in \mathrm{N}_p\} = 1$, owing to the strong law of large numbers. Let $\mu := \mathrm{P} \circ X^{-1}$ denote the distribution of X; we have just seen that μ is a probability measure on N_p.

We begin with a direct computation: If $x \in \mathrm{N}_p$ is fixed, then

$$\begin{aligned} \mathrm{P}\left\{X_1 = x_1\,, \ldots, X_n = x_n\right\} &= p^{nf(x\,;n)}(1-p)^{n-f(x\,;n)} \\ &= 2^{-n(H(p)+o(1))} \qquad \text{as } n \to \infty, \end{aligned} \tag{2.5}$$

where $f(x\,;n) := \sum_{j=1}^{n} \mathbf{1}_{\{0\}}(x)$. Note that the little-$o$ term in (2.5) is allowed to depend on the point x.

Consider the dyadic cube $U_n(x) := \{y \in (0\,,1] :\ y_1 = x_1\,, \ldots, y_n = x_n\}$. The preceding shows that $\mu(U_n(x)) = 2^{-n\{H(p)+o(1)\}}$ for all $x \in \mathrm{N}_p$. Since $y \in U_n(x)$ if and only if $x \in U_n(y)$, it follows fairly easily that $\mu(U_n(x)) = 2^{-n\{H(p)+o(1)\}}$ for all $x \in (0\,,1]$. It is possible to prove that the following hold:

1. $U_n(x) \subseteq [x - 2^{-n},\, x + 2^{-n}]$; and
2. $[x - 2^{-n-1},\, x + 2^{-n-1}] \subset U_n(y) \cup U_n(z)$ for some $y, z \in (0\,,1]$ that might – or might not – be distinct.

A monotonicity argument then shows that for our $\mu \in \mathcal{P}(\mathrm{N}_p)$,

$$\mu([x-r\,, x+r]) = r^{H(p)+o(1)} \qquad \text{as } r \downarrow 0\text{, for all } x \in (0\,,1]. \tag{2.6}$$

The density theorem of Rogers and Taylor [45] finishes the proof. □

2.3. The range of Brownian motion

Let $\mathbf{B} := \{B(t)\}_{t \geq 0}$ denote Brownian motion in $\mathbf{R}^d$. That is, $\mathbf{B}$ is a collection of random variables that satisfy the following: (a) $B(0) := 0$; (b) $B(t+s) - B(s)$ is independent of $\{B(u)\}_{0 \leq u \leq s}$ for all $s, t \geq 0$; and (c) the coordinates of the random vector $B(t+s) - B(s)$ are independent mean-zero gaussian random variables with variance t, regardless of the value of $s \geq 0$ and $t > 0$.

A well-known theorem of Wiener states that one can construct $\mathbf{B}$ such that the resulting random function $t \mapsto B(t)$ is almost surely Hölder continuous with any given index $< 1/2$. In fact, the following limit exists with probability one:

$$\lim_{\epsilon \to 0} \sup_{\substack{s \in [0,T] \\ t \in (s, s+\epsilon]}} \frac{|B(t) - B(s)|}{\sqrt{2(t-s)\ln(t-s)}} = 1 \qquad \text{for all } T > 0. \tag{2.7}$$

This is due to Lévy [31] but with a lim sup in place of the limit; the present, more elegant, formulation can be found in Orey and Pruitt [39] and Csörgő and Révész [7].

Theorem 2.3 (Lévy [31], **Taylor** [51]**).** $\dim_{\mathrm{H}} B[0\,,b] = \min(d\,,2)$ *almost surely for all* $b > 0$.

Proof. We need to prove only the lower bound; the upper bound follows readily from the Hölder continuity of W and (1.7).

Let us define a Borel measure μ via

$$\mu(V) := \int_0^b \mathbf{1}_{\{B(s)\in V\}}\,\mathrm{d}s. \tag{2.8}$$

I plan to prove that $\mathrm{E}(I_q(\mu)) < \infty$ for every positive $q < \min(d\,,2)$.

Note that

$$\mathrm{E}\left(I_q(\mu)\right) = \mathrm{E}\left(\int_{[0,b]^2} \frac{\mathrm{d}s\,\mathrm{d}t}{\|B(s)-B(t)\|^q}\right). \tag{2.9}$$

Elementary properties of gaussian random vectors show that the distribution of $B(t)-B(s)$ is the same as the distribution of $|t-s|^{1/2}$ times $B(1)$, and hence

$$\mathrm{E}\left(I_q(\mu)\right) = \int_{[0,b]^2} \frac{\mathrm{d}s\,\mathrm{d}t}{|s-t|^{q/2}} \cdot \mathrm{E}\left(\|B(1)\|^{-q}\right). \tag{2.10}$$

Since the $(\mathrm{d}s \times \mathrm{d}t)$-integral is finite iff $q<2$, it suffices to prove that $\mathrm{E}(\|B(1)\|^{-q})$ is finite when $q<d$. But direct computation reveals that

$$\mathrm{E}\left(\|B(1)\|^{-q}\right) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \|z\|^{-q}\,\mathrm{e}^{-\|z\|^2/2}\,\mathrm{d}z, \tag{2.11}$$

and the latter integral is finite iff $q<d$. The theorem follows. □

2.4. Fractal percolation

Mandelbrot [34] has introduced the following random Cantor set in the context of turbulence:[4] Choose and fix a parameter $p \in (0\,,1)$, and let $\{Z(I)\}_{I\in\mathcal{D}}$ be an independent, identically-distributed collection of random variables, each taking the values zero and one with respective probabilities $1-p$ and p. Define

$$Q_n(p) := \bigcup \left\{ I \in \mathcal{D}_n \cap (0\,,1]^d :\; Z(I)=1 \right\} \qquad \text{for all } n \geq 0. \tag{2.12}$$

And

$$\Lambda_n(p) := \bigcap_{j=0}^{n} Q_j(p) \qquad \text{for all } n \geq 0. \tag{2.13}$$

In words, in order to construct $\Lambda_n(p)$ from $\Lambda_{n-1}(p)$, we consider each dyadic $I \subset \Lambda_{n-1}(p)$ of sidelength 2^{-n} and retain it in $\Lambda_n(p)$ – independently of the others – with probability p. Since $\Lambda_n(p) \subseteq \Lambda_{n-1}(p)$, the following “fractal-percolation set” is well defined:

$$\Lambda_p := \bigcap_{n=0}^{\infty} \Lambda_n(p). \tag{2.14}$$

Let N_n denote the number of all $I \in \mathcal{D}_n \cap (0\,,1]^d$ that are in $\Lambda_n(p)$. Then, a little thought shows that $\{N_n\}_{n=0}^{\infty}$ is a Galton–Watson branching process with mean offspring distribution $p2^d$. Therefore, the theory of branching processes tells

[4] Mandelbrot refers to this as “random curdling.”

us the following: *If* $p > 2^{-d}$ *then* $\sigma := \mathrm{P}\{\Lambda_p \neq \varnothing\} > 0$*; else, if* $p \le 2^{-d}$ *then* $\sigma = 0$.[5] Thus, from now on we concentrate on the nontrivial case that

$$p > 2^{-d}. \tag{2.15}$$

Note that $\Lambda_n(p)$ is a disjoint union of N_n dyadic cubes of sidelength 2^{-n} each. Therefore, its Lebesgue measure is $N_n 2^{-nd}$. A standard computation reveals that $\mathrm{E}(N_n) = p^n$; see (2.17) below, for example. Therefore, Λ_p has zero Lebesgue measure, almost surely.

Theorem 2.4 (Falconer [13]**, Mauldin and Williams** [36]**; see also Grimmett** [18]**).** *If* (2.15) *holds, then*

$$\|\dim_{\mathrm{H}} \Lambda_p\|_{L^\infty(\mathrm{P})} = \left\|\overline{\dim}_{\mathrm{M}} \Lambda_p\right\|_{L^\infty(\mathrm{P})} = d - \log_2(1/p), \tag{2.16}$$

where "$\log_2$*" denotes the logarithm in base* 2*.*

Proof. Choose and fix two integers $n > k \ge 1$. The probability that a given dyadic interval $I \in \mathcal{D}_k \cap (0\,,1]$ intersects $\Lambda_n(p)$ is p^k, since I intersects $\Lambda_n(p)$ if and only if $Z(J) = 1$ for all k dyadic cubes that include I. Consequently,

$$\mathrm{P}\left\{I \cap \Lambda_p \neq \varnothing\right\} \le p^k. \tag{2.17}$$

Let N_k denote the number of dyadic cubes $I \in \mathcal{D}_k$ that intersect Λ_p. We can sum (2.17) over all $I \in \mathcal{D}_k$ to find that $\mathrm{E}(N_k) \le (2^d p)^k$. Therefore, Chebyshev's inequality implies that for every positive $\lambda < 2^d p$,

$$\sum_{k=1}^\infty \mathrm{P}\left\{N_k > \lambda^k\right\} \le \sum_{k=1}^\infty \frac{(2^d p)^k}{\lambda^k} < \infty. \tag{2.18}$$

Since $\lambda \in (0\,, 2^d p)$ is arbitrary, the preceding and the Borel–Cantelli lemma together show that $\overline{\dim}_{\mathrm{M}} \Lambda_p \le \log_2(2^d p) = d - \log_2(1/p)$ almost surely.

In order to prove the remainder of the theorem, we apply an elegant "intersection argument" of Peres [41]. Let G be a nonrandom Borel set in $(0\,,1]^d$. I will prove the following:

Claim A. *If* $\dim_{\mathrm{H}} G < \log_2(1/p)$*, then* $\Lambda_p \cap G = \varnothing$ *almost surely.*

We first use this to prove Theorem 2.4, and then establish Claim A.

Consider an independent fractal percolation Λ'_β. Because $\Lambda_p \cap \Lambda'_\beta$ has the same distribution as $\Lambda_{p\beta}$, we see that if $p\beta 2^d > 1$ then $\Lambda_p \cap \Lambda'_\beta \neq \varnothing$ with positive probability. We condition on Λ_p to deduce from Claim A that $\dim_{\mathrm{H}} \Lambda_p \ge \log_2(1/\beta)$ with positive probability. Since any positive $1/\beta < p2^d$ leads to this bound, it follows that $\|\dim_{\mathrm{H}} \Lambda_p\|_{L^\infty(\mathrm{P})} \ge \log_2(2^d p) = d - \log_2(1/p)$, as needed. Now let us prove Claim A.

[5] In fact, if we define $\mathcal{N} := \#\{I \in \mathcal{D}_1 :\ Z(I) = 1\}$ – that is, $\mathcal{N}$ denotes the number of retained level-1 dyadic cubes – then $\sigma = \mathrm{E}(\sigma^{\mathcal{N}})$ by independence. Since $\mathcal{N}$ has a binomial distribution with parameters 2^d and p, the assertion follows from direct computation.

We fix an arbitrary $\epsilon > 0$, and find dyadic cubes $U_1, U_2, \ldots$ such that: (i) $\{U_j\}_{j=1}^{\infty}$ is a cover for G; and (ii) $\sum_{j=1}^{\infty} |\text{side } U_j|^{\log_2(1/p)} < \epsilon$. In accord with (2.17),

$$\mathrm{P}\left\{\Lambda_p \cap G \neq \varnothing\right\} \le \sum_{j=1}^{\infty} \mathrm{P}\left\{\Lambda_p \cap U_j \neq \varnothing\right\} = \sum_{j=1}^{\infty} p^{\log_2(1/\text{side } U_j)} < \epsilon. \tag{2.19}$$

This completes our proof, since $\epsilon > 0$ is arbitrary. □

Remark 2.5 (Theorem 2.4, revisited). Because Λ_p has a great deal of self-similarity, Theorem 2.4 can be improved to the following [13, 36], which we prove next:

$$\mathrm{P}\left\{\dim_{\mathrm{H}} \Lambda_p = \overline{\dim}_{\mathrm{M}} \Lambda_p = d - \log_2(1/p) \;\middle|\; \Lambda_p \neq \varnothing\right\} = 1. \tag{2.20}$$

Let us recall (2.15), let $\mathcal{N}$ denote the number of $I \in \mathcal{D}_1$ such that $Z(I) = 1$, and choose $\lambda \in (0\,, d - \log_2(1/p))$. We can consider the probabilities

$$\alpha(I) := \mathrm{P}\{\dim_{\mathrm{H}}(\Lambda_p \cap I) \le \lambda\ ,\ \Lambda_p \cap I \neq \varnothing\}, \tag{2.21}$$

defined for all dyadic cubes $I \subset (0\,,1]^d$. We can condition on $\{Z(I)\}_{I \in \mathcal{D}_1}$ to find that $\alpha((0\,,1]^d) \le \mathrm{E}(\prod \alpha(I)\ ;\ \mathcal{N} \ge 1)$, where the product is over all $I \in \mathcal{D}_1 \cap (0\,,1]^d$ such that $Z(I) = 1$. But $\alpha(I) = \alpha((0\,,1]^d)$ for every dyadic cube $I \subseteq (0\,,1]^d$. Therefore,

$$\begin{aligned} \alpha\left((0\,,1]^d\right) &\le \mathrm{E}\left(\left[\alpha\left(0\,,1]^d\right)\right]^{\mathcal{N}}\ ;\ \mathcal{N} \ge 1\right) \\ &\le \alpha\left((0\,,1]^d\right) \cdot \mathrm{P}\{\mathcal{N} \neq 0\}. \end{aligned} \tag{2.22}$$

Theorem 2.4 implies that $\alpha((0\,,1]^d) < 1$. Since $\mathrm{P}\{\mathcal{N} \neq 0\} < 1$, the preceding proves that $\alpha((0\,,1]^d) = 0$, which implies (2.20) readily. □

3. Lévy processes

Lévy processes are a natural family of random processes that "cannot help but produce fractals." I present a very brief introduction to the general theory. The books by Bertoin [3], Jacob [23], and Sato [48] are excellent, and thorough, accounts of Lévy processes.

A *Lévy process* $\mathbf{X} := \{X(t)\}_{t \ge 0}$ on $\mathbf{R}^d$ is a stochastic process [that is, a collection of random variables] which satisfies the following properties:

1. $X(0) = 0$ almost surely;
2. $X(t+s) - X(s)$ is independent of $\{X(u)\}_{0 \le u \le s}$ for all $s, t \ge 0$;
3. The distribution of $X(t+s) - X(s)$ does not depend on s, for all $s, t \ge 0$; and
4. $t \mapsto X(t)$ is continuous in probability; i.e., $\lim_{s \to t} \mathrm{P}\{|X(s) - X(t)| > \epsilon\} = 0$ for all $t \ge 0$ and $\epsilon > 0$.

It turns out that one can always arrange things such that $t \mapsto X(t)$ is right-continuous and has left limits at all points [3, p. 13]; in particular, $\mathbf{X}$ can have only discontinuities of the first kind, if it is at all discontinuous.

According to the Lévy–Khintchine formula (Bertoin [3, Theorem 1, p. 13]), a stochastic process $\mathbf{X}$ is a Lévy process if and only if

$$\mathrm{E}\left(\mathrm{e}^{i\xi\cdot X(t)}\right) = \mathrm{e}^{-t\Psi(\xi)} \qquad \text{for all } \xi \in \mathbf{R}^d \text{ and } t > 0, \tag{3.1}$$

where Ψ is a negative-definite function – in the sense of Schoenberg [49] – such that $\Psi(0) = 0$; see also the monographs by Jacob [23] and Sato [48]. An equivalent statement is this: If $\mu_t(A) := \mathrm{P}\{X(t) \in A\}$ defines the distribution of $X(t)$, then $\{\mu_t\}_{t\geq 0}$ is a weakly-continuous convolution-semigroup of probability measures with $\mu_0 := \delta_0$ [3, Chapter 1]. Of course, in this case we have $\int_{\mathbf{R}^d} \exp(i\xi \cdot x)\mu_t(\mathrm{d}x) = \exp(-t\Psi(\xi))$. In general, we might refer to the function Ψ as the *characteristic exponent*, or *Lévy exponent*, of the process $\mathbf{X}$.

Example. It might be good to keep some examples of Lévy processes in mind:

1. If $\Psi(\xi) = c\|\xi\|^\alpha$ for some $c > 0$, all $\xi \in \mathbf{R}^d$, and $\alpha \in (0\,,2]$, then $\mathbf{X}$ is an isotropic stable process on $\mathbf{R}^d$ with index α. When $\alpha = 2$ and $c = 1/2$, $\mathbf{X}$ is a Brownian motion.
2. The *Poisson process* is a well-known Lévy process on $\mathbf{Z}_+ := \{0\,,1\,,\ldots\}$; its characteristic exponent is $\Psi(\xi) = \lambda(1 - \mathrm{e}^{i\xi})$ for $\xi \in \mathbf{R}$ and $\lambda > 0$ is its *rate.*
3. A *compensated Poisson process* is a Lévy process on $\mathbf{R}$, and its characteristic exponent is $\Psi(\xi) = \lambda(1+i\xi-\mathrm{e}^{i\xi})$. If $\mathbf{Y}$ is a rate-λ Poisson process on $\mathbf{Z}_+$, then $X(t) := Y(t) - \lambda t$ defines a compensated Poisson process with rate λ. □

Perhaps one of the most common features of many interesting fractals is that they have zero Lebesgue measure. The next result is a characterization of all Lévy processes whose range has zero Lebesgue measure; those are Lévy processes that tend to "generate" random fractals. With this in mind, let us define

$$\kappa(\xi) := \mathrm{Re}\left(\frac{1}{1+\Psi(\xi)}\right) \qquad \text{for all } \xi \in \mathbf{R}^d. \tag{3.2}$$

Theorem 3.1 (Kesten [26]**; see also Orey** [38]**).** *Choose and fix $b > 0$. Then the Lebesgue measure of $X[0\,,b]$ is positive with positive probability if and only if $\kappa \in L^1(\mathbf{R}^d)$.*

Remark 3.2. It is well known that Brownian motion is "extremal" among all Lévy processes in the sense that $|\Psi(\xi)| \leq \mathrm{const}\cdot(1+\|\xi\|^2)$. Therefore, $\kappa(\xi) \geq \mathrm{const}/(1+\|\xi\|^2)$; see Bochner [4, eq. (3.4.14), p. 67], for instance. As a result, we can deduce the following theorem of Lévy [31, Théorèm 53, p. 256]: *If $d \geq 2$ then $X[0\,,b]$ has zero Lebesgue measure almost surely.* □

Remark 3.3. The passing reference to fractals should not, and can not, be taken too seriously. For example, if $\mathbf{X}$ is a Poisson process on the line with rate $\lambda > 0$, then $\Psi(\xi) = \lambda(1-\mathrm{e}^{i\xi})$, and $\kappa \notin L^1(\mathbf{R})$. Thus, Theorem 3.1 verifies the elementary fact that the range of $\mathbf{X}$ has zero Lebesgue measure. That range is $\mathbf{Z}_+$, which is clearly not an interesting fractal. □

It is more difficult to compute the Hausdorff dimension of $X[0\,,b]$ than to decide when $X[0\,,b]$ has positive Lebesgue measure. In order to describe the Hausdorff dimension of $X[0\,,b]$ we consider the "Cauchy transform" W of κ,

$$W(r) := \int_{\mathbf{R}^d} \frac{\kappa(\xi/r)}{\prod_{j=1}^d (1+\xi_j^2)}\,\mathrm{d}\xi \qquad \text{for all } r > 0. \tag{3.3}$$

Theorem 3.4 (Khoshnevisan and Xiao [28]**).** *For all $b > 0$, the following holds with probability one:*

$$\dim_{\mathrm{H}} X[0\,,b] = \underline{\dim}_{\mathrm{M}} X[0\,,b] = \liminf_{r\to 0} \frac{\log W(r)}{\log r}. \tag{3.4}$$

Remark 3.5. One can also prove that almost surely,

$$\dim_{\mathrm{P}} X[0\,,b] = \overline{\dim}_{\mathrm{M}} X[0\,,b] = \limsup_{r\to 0} \frac{\log W(r)}{\log r}, \tag{3.5}$$

where $\dim_{\mathrm{P}}$ denotes the packing dimension [28]. □

Example. It is possible to check directly that if $\mathbf{X}$ is Brownian motion on $\mathbf{R}^d$, then $\dim_{\mathrm{H}} X[0\,,b] = \min(d\,,2)$ almost surely. This agrees with Theorem 2.3. □

3.1. Subordinators: An example

A real-valued Lévy process $\mathbf{X}$ is a *subordinator* if $t \mapsto X(t)$ is almost surely increasing and everywhere nonnegative. We have seen already that one can characterize a subordinator $\mathbf{X}$ by its characteristic exponent Ψ. But it is sometimes simpler to consider its *Laplace exponent* $\Phi : \mathbf{R}_+ \to \mathbf{R}_+$ [2]; the defining feature of Φ is that it solves $\mathrm{E}\exp(-\xi X(t)) = \exp(-t\Phi(\xi))$ for all $t, \xi \geq 0$. There are various relationships between the Laplace exponent Φ and the characteristic exponent Ψ. We mention one next: Let $\mathbf{S} := \{S(t)\}_{t\geq 0}$ denote an independent symmetric Cauchy process [that is, $\mathbf{S}$ is a Lévy process with characteristic exponent $\Psi(\xi) := |\xi|$, so that $S(t)/t$ has the standard Cauchy distribution on the line for all $t > 0$] and note that the following from a few applications of Fubini's theorem: For all $t, \xi \geq 0$,

$$\mathrm{e}^{-t\Phi(\xi)} = \mathrm{E}\left[\mathrm{e}^{-\xi X(t)}\right] = \mathrm{E}\left[\mathrm{e}^{iX(t)S(\xi)}\right] = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-t\Psi(\xi z)}}{1+z^2}\,\mathrm{d}z. \tag{3.6}$$

We integrate both side $[\mathrm{e}^{-t}\,\mathrm{d}t]$ to find that

$$\frac{1}{1+\Phi(\xi)} = \frac{1}{\pi} W(1/\xi) \qquad \text{for all } \xi \geq 0. \tag{3.7}$$

Theorem 3.6 implies the following theorem of Horowitz [22]: With probability one,

$$\dim_{\mathrm{H}} X[0\,,b] = \limsup_{\xi\to\infty} \frac{\log \Phi(\xi)}{\log \xi}. \tag{3.8}$$

For an example, let us consider a one-dimensional Brownian motion $\mathbf{B}$; the theory of Brownian local times tells us that the zero set of $\mathbf{B}$ is the closure of the range of a subordinator $\mathbf{X}$ with $\Phi(\xi) = \text{const}\cdot \xi^{1/2}$ for all positive ξ; see Maisonneuve [33], and also Bertoin [2, Chapter 9] for a pedagogic account. Because the closure of the range of $\mathbf{X}$ and the range itself differ in at most the jump points of $\mathbf{X}$ – and there are only countably-many of those – we can conclude the following well-known theorem of Lévy [31]: *The Hausdorff dimension of the zero-set of Brownian motion is almost surely equal to* $\limsup_{\xi\to\infty} \log\Phi(\xi)/\log\xi = 1/2$. Lévy's theorem [for Brownian motion] has simpler proofs than the one outlined here. But the present method can be used to compute the Hausdorff dimension of the level sets of quite general Markov processes, when simpler arguments no longer exist.

3.2. Proof of Theorem 3.1, and the odds of hitting a ball

Define, for each $b \geq 0$, the *incomplete renewal measure* U_b as follows:

$$U_b(A) := \int_0^b \mathrm{P}\{X(s) \in A\}\,\mathrm{d}s. \tag{3.9}$$

Each U_b is a finite Borel measure on $\mathbf{R}^d$ of total mass b.

Define the ball

$$B(x\,,r) := \left\{y \in \mathbf{R}^d :\ |x-y| \leq r\right\}, \tag{3.10}$$

where

$$|z| := \max_{1\leq j\leq d} |z_j| \qquad \text{for all } z \in \mathbf{R}^d. \tag{3.11}$$

Then we have the following "quantitative hitting-time" estimate. The particular formulation that follows appeared in [27], but this is an old folklore result which arises in various forms in many parts of the literature.

Theorem 3.6. *The following holds for all* $x \in \mathbf{R}^d$ *and* $b, r > 0$*:*

$$\frac{U_b(B(x\,,r))}{U_b(B(0\,,2r))} \leq \mathrm{P}\left\{X[0\,,b] \cap B(x\,,r) \neq \varnothing\right\} \leq \frac{U_{2b}(B(x\,,2r))}{U_b(B(0\,,r))}. \tag{3.12}$$

Proof. Let T denote the smallest time $s \in [0\,,b]$ at which $|X(s)-x| \leq r$; if such an s does not exist, then set $T := b+1$. We can write

$$\begin{aligned} U_b(B(x\,,r)) &= \mathrm{E}\left(\int_T^b \mathbf{1}_{\{|X(s)-x|\leq r\}}\,\mathrm{d}s\ ;\ 0 \leq T \leq b\right) \\ &= \mathrm{E}\left(\int_0^{b-T} \mathbf{1}_{\{|X(s+T)-X(T)+X(T)-x|\leq r\}}\,\mathrm{d}s\ ;\ 0 \leq T \leq b\right) \\ &\leq \mathrm{E}\left(\int_0^{b} \mathbf{1}_{\{|X(s+T)-X(T)+X(T)-x|\leq r\}}\,\mathrm{d}s\ ;\ 0 \leq T \leq b\right). \end{aligned} \tag{3.13}$$

According to the strong Markov property [3, Proposition 6, p. 20], the process $\{X(s+T)-X(T)\}_{s\geq 0}$ is a copy of $\mathbf{X}$, and is independent of $\{X(u)\}_{u\in[0,T]}$. It follows from this that

$$\begin{aligned} U_b(B(x\,,r)) &\leq \mathrm{E}\left[U_b(B(X(T)-x\,;r))\ ;\ 0 \leq T \leq b\right] \\ &\leq U_b(B(0\,,2r)) \cdot \mathrm{P}\{0 \leq T \leq b\}, \end{aligned} \tag{3.14}$$

because if $T \in [0\,,b]$ then $|X(T)-x| \le r$, whence $B(X(T)-x\,,r) \subseteq B(0\,,2r)$. This proves the first inequality of the theorem. The second inequality is proved similarly, but instead of the preceding with start with the following: For the same T as before,

$$\begin{aligned}
U_{2b}(B(x\,,2r)) &\ge \mathrm{E}\left(\int_T^{2b} \mathbf{1}_{\{|X(s)-x|\le 2r\}}\,\mathrm{d}s\ ;\ 0 \le T \le b\right)\\
&= \mathrm{E}\left(\int_0^{2b-T} \mathbf{1}_{\{X(T+s)-X(T)+X(T)-x|\le 2r\}}\,\mathrm{d}s\ ;\ 0 \le T \le b\right)\\
&\ge \mathrm{E}\left(\int_0^{b} \mathbf{1}_{\{X(T+s)-X(T)+X(T)-x|\le 2r\}}\,\mathrm{d}s\ ;\ 0 \le T \le b\right)\\
&= \mathrm{E}\left(U_b(B(X(T)-x\,,2r))\ ;\ 0 \le T \le b\right).
\end{aligned} \tag{3.15}$$

The theorem follows from this and an application of the triangle inequality; namely, that $B(X(T)-x\,,2r) \supseteq B(0\,,r)$ almost surely on $\{0 \le T \le b\}$. □

Recall that the distribution of $X(s)$ is μ_s. Thus, we define the *renewal measure* U of the process $\mathbf{X}$ via

$$U(A) := \int_0^\infty \mathrm{P}\{X(s) \in A\}\,\mathrm{e}^{-s}\,\mathrm{d}s = \int_0^\infty \mu_s(A)\mathrm{e}^{-s}\,\mathrm{d}s. \tag{3.16}$$

Note that U is a Borel probability measure on $\mathbf{R}^d$. Next we show that the complete renewal measure is estimated well by the incomplete ones.

Lemma 3.7. *For all $b, r > 0$,*

$$\mathrm{e}^{-b}U_b(B(0\,,r)) \le U(B(0\,,r)) \le \frac{16^d}{1-\mathrm{e}^{-b}} \cdot U_b(B(0\,,r)). \tag{3.17}$$

Proof. The first inequality is an elementary consequence of the definitions of U and U_b; we derive the second one only.

We can write

$$\begin{aligned}
U(B(0\,,r)) &\le \sum_{k=0}^\infty \mathrm{e}^{-kb} \int_{kb}^{(k+1)b} \mathrm{P}\{|X(s)| \le r\}\,\mathrm{d}s\\
&= \sum_{k=0}^\infty \mathrm{e}^{-kb} \int_0^b \mathrm{P}\{|X(s+kb)-X(kb)+X(kb)| \le r\}\,\mathrm{d}s\\
&= \sum_{k=0}^\infty \mathrm{e}^{-kb}\mathrm{E}\left[U_b\left(B(X(kb)\,,r)\right)\right],
\end{aligned} \tag{3.18}$$

since $\{X(s+kb)-X(kb)\}_{s\ge 0}$ has the same distribution as the Lévy process $\mathbf{X}$, and is also independent of $X(kb)$.

Since probabilities are ≤ 1, the first inequality in Theorem 3.6 tells us that

$$\sup_{x\in\mathbf{R}^d} U_b(B(x\,,r)) \le U_b(B(0\,,2r)) \qquad \text{for all } r > 0. \tag{3.19}$$

Because we can cover $B(0\,,2r)$ by at most 16^d disjoint balls of radius $(r/2)$, the preceding shows that

$$U_b(B(0\,,2r)) \le 16^d U(B(0\,,r)). \tag{3.20}$$

One more application of (3.19) shows that

$$\sup_{x\in\mathbf{R}^d} U_b(B(x\,,r)) \le 16^d U_b(B(0\,,r)) \qquad \text{for all } r>0. \tag{3.21}$$

This and (3.18) together imply the second inequality of the lemma. □

Proof of Theorem 3.1. Let $R(r)$ denote the r-enlargement of $-X[0\,,b]$ for every $r>0$. That is, $R(r) := \cup B(-X(s)\,,r)$, where the union is taken over all $s\in[0\,,b]$. We might note that $\mathrm{P}\{X[0\,,b]\cap B(x\,,r)\neq\varnothing\} = \mathrm{E}[\mathbf{1}_{R(r)}(x)]$, and hence

$$\int_{\mathbf{R}^d} \mathrm{P}\left\{X[0\,,b]\cap B(x\,,r)\neq\varnothing\right\}\,\mathrm{d}x = \mathrm{E}\left[\mathrm{leb}(R(r))\right], \tag{3.22}$$

where "leb" denotes the Lebesgue measure on $\mathbf{R}^d$. We can integrate the inequalities of Theorem 3.6 with respect to $\mathrm{d}x$ to find that

$$\mathrm{E}\left[\mathrm{leb}(R(r))\right] \asymp \frac{r^d}{U_b(B(0\,,r))}, \tag{3.23}$$

where $f\asymp g$ means that $(f/g)(r)$ is bounded away from zero and infinity by constants, uniformly over all $r>0$. [The preceding requires the Tonnelli theorem and the fact that the total mass of U_b is finite and positive.]

Clearly, $\mathrm{leb}(R(r))$ converges downward to the Lebesgue measure of $X[0\,,b]$ as $r\downarrow 0$. Therefore, it suffices to prove that κ is integrable iff $U_b(B(0\,,r)) = O(r^d)$ as $r\downarrow 0$. Thanks to (3.20), it remains to prove the following:

$$\kappa\in L^1(\mathbf{R}^d) \qquad \text{iff} \qquad U(B(0\,,r)) = O(r^d) \text{ as } r\downarrow 0. \tag{3.24}$$

We use Fourier analysis to establish this, and hence the theorem.

Owing to (3.16), $\hat{U}(\xi) = \{1+\Psi(\xi)\}^{-1}$ defines the Fourier transform of U in the sense of distributions; in particular, $|\hat{U}(\xi)|\le 1$. Consequently, for all rapidly-decreasing test functions $\phi:\mathbf{R}^d\to\mathbf{R}$,

$$\int \phi\,\mathrm{d}U = \frac{1}{(2\pi)^d}\int_{\mathbf{R}^d} \mathrm{Re}\left(\frac{\overline{\hat{\phi}(\xi)}}{1+\Psi(\xi)}\right)\mathrm{d}\xi. \tag{3.25}$$

Next we prove that the preceding holds for all uniformly continuous ϕ such that $\hat{\phi}$ is real and nonnegative. Indeed, let γ_n denote the density function of the centered gaussian distribution on $\mathbf{R}^d$ whose covariance matrix is $1/n$ times the identity, and then apply (3.25) to $\phi*\gamma_n$ in place of ϕ to find that

$$\int (\phi*\gamma_n)\,\mathrm{d}U = \frac{1}{(2\pi)^d}\int_{\mathbf{R}^d} \mathrm{Re}\left(\frac{\overline{\hat{\phi}(\xi)}}{1+\Psi(\xi)}\right)\mathrm{e}^{-\|\xi\|^2/(2n)}\,\mathrm{d}\xi, \tag{3.26}$$

for all $n\ge 1$. Because ϕ is uniformly continuous, $\phi*\gamma_n\to\phi$ uniformly as $n\to\infty$; and the left-hand side of (3.26) converges to $\int\phi\,\mathrm{d}U$ as $n\to\infty$. Because $\hat{\phi}\ge 0$, (3.25) follows from applying the monotone convergence theorem to the right-hand side of (3.26).

Let $f_r(x) := (2r)^{-d}\mathbf{1}_{B(0,r)}(x)$ and $\phi := \phi_r := f_r * f_r$. Since ϕ_r is uniformly continuous and $\hat{\phi}_r = |\hat{f}_r|^2 \geq 0$, it follows that

$$\int \phi_r \, \mathrm{d}U = \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \hat{\phi}_r(\xi)\kappa(\xi)\,\mathrm{d}\xi. \tag{3.27}$$

Because $2^{-d}(2r)^{-d}\mathbf{1}_{B(0,r/2)} \leq \phi_r \leq (2r)^{-d}\mathbf{1}_{B(0,2r)}$,

$$\frac{U(B(0\,,r/2))}{2^d \cdot (2r)^d} \leq \int \phi_r \, \mathrm{d}U \leq \frac{U(B(0\,,2r)}{(2r)^d}. \tag{3.28}$$

Therefore, (3.24) is equivalent to the statement that $\kappa \in L^1(\mathbf{R}^d)$ iff $\int \phi_r \, \mathrm{d}U = O(1)$. Suppose, first, that κ is integrable. Since $|\hat{\phi}_r| \leq 1$, (3.27) implies that $\limsup_{r\downarrow 0} \int \phi_r \, \mathrm{d}U \leq (2\pi)^{-d}\|\kappa\|_{L^1(\mathbf{R}^d)} < \infty$. For the converse, we merely observe that $(2\pi)^{-d}\|\kappa\|_{L^1(\mathbf{R}^d)} \leq \liminf_{r\downarrow 0} \int \phi_r \, \mathrm{d}U$, thanks to Fatou's lemma and the fact that $\lim_{r\downarrow 0} \hat{\phi}_r(\xi) = 1$ for all ξ. □

3.3. A theorem of Pruitt, and proof of Theorem 3.4

Theorem 3.8 (Pruitt [44]). *For all $b > 0$, the following holds almost surely:*

$$\dim_{\mathrm{H}} X[0\,,b] = \underline{\dim}_{\mathrm{M}} X[0\,,b] = \liminf_{r\downarrow 0} \frac{\log U\left(B(0\,,r)\right)}{\log r} \tag{3.29}$$

Proof. Let $N_n(b)$ denote the total number of dyadic cubes $I \in \mathcal{D}_n$ that intersect $X[0\,,b]$. According to Theorem 3.6,

$$\mathrm{E}(N_n(b)) \leq \frac{1}{U_b(B(0\,,2^{-n-1}))} \cdot \sum_{x:\ B(x,2^{-n-1}) \in \mathcal{D}_n} U_{2b}\left(B(x\,,2^{-n})\right). \tag{3.30}$$

Because of (3.19) and (3.20),

$$\sup_{n\geq 1} \sum_{x:\ B(x,2^{-n-1}) \in \mathcal{D}_n} U_{2b}\left(B(x\,,2^{-n})\right) < \infty. \tag{3.31}$$

Consequently,

$$\mathrm{E}(N_n(b)) \leq \frac{\mathrm{const}}{U(B(0\,,2^{-n}))}, \tag{3.32}$$

thanks to Lemma 3.7 and (3.18). Let us select $q > 0$ such that $U(B(0\,,2^{-n})) \geq 2^{-nq}$ for infinitely-many n tending to ∞. Then, (3.32) and Fatou's lemma together imply that $\liminf_{n\to\infty} 2^{-nq}N_n(b) < \infty$ almost surely. Therefore, $\dim_{\mathrm{H}} X[0\,,b] \leq \underline{\dim}_{\mathrm{M}} X[0\,,b] \leq q$ almost surely; this proves half of the theorem.

We prove the other half by appealing to Frostman's theorem (Theorem 1.1). Consider the following [random] Borel measure μ [compare with (2.8)]:

$$\mu(V) := \int_0^b \mathbf{1}_V(X(s))\,\mathrm{d}s. \tag{3.33}$$

Then, $I_q(\mu) = \int_{[0,b]^2} |X(s) - X(t)|^{-q}\,\mathrm{d}s\,\mathrm{d}t$ for all $q > 0$. We compute expectations.

$$\mathrm{E}\left(I_q(\mu)\right) = 2 \iint\limits_{0\le s<t<b} \mathrm{E}\left(|X(t-s)|^{-q}\right)\,\mathrm{d}s\,\mathrm{d}t \le 2b \int_0^b \mathrm{E}\left(|X(s)|^{-q}\right)\,\mathrm{d}s; \tag{3.34}$$

then we integrate by parts to find that

$$\mathrm{E}\left(|X(s)|^{-q}\right) = q \int_0^\infty \lambda^{-q-1} \mathrm{P}\left\{|X(s)| \le \lambda\right\}\,\mathrm{d}\lambda. \tag{3.35}$$

It follows from the preceding and Lemma 3.7 that

$$\begin{aligned}\mathrm{E}\left(I_q(\mu)\right) &\le 2bq \int_0^\infty \lambda^{-q-1} U_b\left(B(0\,,\lambda)\right)\,\mathrm{d}\lambda \\ &\le 2bq\mathrm{e}^b \cdot \int_0^\infty \lambda^{-q-1} U\left(B(0\,,\lambda)\right)\,\mathrm{d}\lambda.\end{aligned} \tag{3.36}$$

If $q > 0$ and $\epsilon > 0$ are selected so that $U(B(0\,,r)) = O(r^{q+\epsilon})$ as $r \to 0$, then $I_q(\mu)$ is almost surely finite, since it will have a finite expectation. This proves the theorem. □

Proof of Theorem 3.4. If $u \in L^1_{\mathrm{loc}}(\mathbf{R}^d)$, or u is a locally finite Borel measure on $\mathbf{R}^d$, then $\hat{u}$ denotes its Fourier transform in the sense of Schwartz. We normalize the Fourier transform so that $\hat{u}(\xi) = \int_{\mathbf{R}^d} u(x)\exp(i\xi\cdot x)\,\mathrm{d}x$ whenever $u \in L^1(\mathbf{R}^d)$.

Consider the function

$$\varphi_r(x) := \prod_{j=1}^d \left(\frac{1-\cos(2rx_j)}{2\pi r x_j^2}\right) \qquad \text{for } x \in \mathbf{R}^d, \tag{3.37}$$

where $r > 0$ is a parameter. Then φ_r is a nonnegative integrable function on $\mathbf{R}^d$, and its Fourier transform is the *Pólya kernel*,

$$\hat{\varphi}_r(\xi) = \prod_{j=1}^d \left(1 - \frac{|\xi_j|}{2r}\right)^+. \tag{3.38}$$

Note that $\hat{\varphi}_r(\xi) \ge 2^{-d}$ whenever $|\xi| \le r$. That is,

$$\mathbf{1}_{B(0,r)}(\xi) \le 2^d \hat{\varphi}_r(\xi) \qquad \text{for all } r > 0 \text{ and } \xi \in \mathbf{R}^d. \tag{3.39}$$

Therefore, by Fubini's theorem,

$$\mathrm{P}\{X(s) \in B(0\,,r)\} \le 2^d \mathrm{E}\left(\hat{\varphi}_r(X(s))\right) = 2^d \int_{\mathbf{R}^d} \hat{\varphi}_r(x)\mu_s(\mathrm{d}x), \tag{3.40}$$

where, we recall, $\mu_s := \mathrm{P}\circ X(s)^{-1}$ denotes the distribution of $X(s)$. Consequently,

$$\mathrm{P}\{X(s) \in B(0\,,r)\} \le 2^d \int_{\mathbf{R}^d} \varphi_r(\xi)\hat{\mu}_s(\xi)\,\mathrm{d}\xi = 2^d \int_{\mathbf{R}^d} \varphi_r(\xi) e^{-s\Psi(\xi)}\,\mathrm{d}\xi. \tag{3.41}$$

We integrate $[\mathrm{e}^{-s}\mathrm{d}s]$ to find that

$$U(B(0\,,r)) \le 2^d \int_{\mathbf{R}^d} \varphi_r(\xi)\kappa(\xi)\,\mathrm{d}\xi; \tag{3.42}$$

only the real part [part of the definition of κ] enters because $U(B(0\,,r))$ and $\phi_r(\xi)$ are real valued. Since $(1-\cos z)/z^2 \leq \text{const}/(1+z^2)$ for all real numbers z, this proves that $U(B(0\,,r)) \leq \text{const}\cdot W(r)$, whence half of the theorem.

For the other inequality we choose and fix $\delta \in (0\,,1)$, and note that for every $r>0$ and $z \in \mathbf{R}^d$,

$$\mathbf{1}_{B(0,r)}(z) \geq \exp\left(-\frac{1}{r^{1-\delta}}\sum_{j=1}^d |z_j|\right) - \exp\left(-r^{-1+\delta}\right). \tag{3.43}$$

Plug in $z := X(s)$, take expectations, and then integrate $[\mathrm{e}^{-s}\,\mathrm{d}s]$ to find that

$$U\left(B(0\,,r)\right) \geq \mathrm{E}\left[\int_0^\infty \exp\left(-\frac{1}{r^{1-\delta}}\sum_{j=1}^d |X_j(s)|\right)\mathrm{e}^{-s}\,\mathrm{d}s\right] - \exp\left(-r^{-1+\delta}\right). \tag{3.44}$$

But with probability one,

$$\exp\left(-\frac{1}{r^{1-\delta}}\sum_{j=1}^d |X_j(s)|\right) = \mathrm{E}\left[\exp\left(i\frac{S\cdot X(s)}{r^{1-\delta}}\right)\,\middle|\, X(s)\right], \tag{3.45}$$

where $S := (S_1,\ldots,S_d)$ is a vector of d independent standard-Cauchy random variables; the probability density function of S is $p(\xi) := \pi^{-d}\prod_{j=1}^d(1+\xi_j^2)^{-1}$ at $\xi \in \mathbf{R}^d$. By Fubini's theorem, and after a calculation, we find that

$$\begin{aligned}
U\left(B(0\,,r)\right) &\geq \mathrm{E}\left[\int_0^\infty \exp\left(i\,\frac{S\cdot X(s)}{r^{1-\delta}}\right)\mathrm{e}^{-s}\,\mathrm{d}s\right] - \exp\left(-r^{-1+\delta}\right)\\
&= \int_0^\infty \mathrm{e}^{-s}\mathrm{d}s\int_{\mathbf{R}^d} p(\xi)\,\mathrm{d}\xi\ \mathrm{E}\left[\exp\left(i\,\frac{\xi\cdot X(s)}{r^{1-\delta}}\right)\right] - \exp\left(-r^{-1+\delta}\right)\\
&= \frac{1}{\pi^d}W(r^{1-\delta}) - \exp\left(-r^{-1+\delta}\right).
\end{aligned} \tag{3.46}$$

And this is sufficient to prove the remaining direction of the theorem. □

3.4. Occupation measures, local times, and Hawkes's theorem

Let $\mathbf{X}$ denote a Lévy process on $\mathbf{R}^d$, and define

$$Q(G) := \int_0^\infty \mathbf{1}_G(X(s))\mathrm{e}^{-s}\,\mathrm{d}s, \tag{3.47}$$

for all Borel sets $G \subseteq \mathbf{R}^d$. Evidently, Q is a [random] Borel probability measure on $\mathbf{R}^d$. We follow Geman and Horowitz [17], and say that $\mathbf{X}$ *has square-integrable local times* if Q is absolutely continuous with respect to leb, and its Radon–Nikodým density ℓ satisfies $\ell := \mathrm{d}Q/\mathrm{d}x \in L^2(\mathbf{R}^d)$. The random process $\{\ell(x)\}_{x\in\mathbf{R}^d}$ is then called the *local times* of $\mathbf{X}$. Note that if $\mathbf{X}$ has square-integrable local times, then the following holds: For all nonrandom Borel-measurable functions $f:\mathbf{R}^d \to \mathbf{R}_+$,

$$\int_0^\infty f(X(s))\mathrm{e}^{-s}\,\mathrm{d}s = \int_{\mathbf{R}^d} f(x)\ell(x)\,\mathrm{d}x \qquad \text{almost surely.} \tag{3.48}$$

In words, local times exist iff Q is differentiable. In this way, local times are the most natural "Frostman-like" measures that can be constructed on the range of a given Lévy process. These local times will make a surprising appearance in the following section on stochastic PDEs, as well.

Theorem 3.9 (Hawkes [20]**).** *A Lévy process* $\mathbf{X}$ *has square-integrable local times if and only if the Lebesgue measure of* $X(\mathbf{R}_+)$ *is positive with positive probability. Another equivalent condition is that* $\kappa \in L^1(\mathbf{R}^d)$*.*

Proof. Theorem 3.1 shows the equivalence of the assertion "$\kappa \in L^1(\mathbf{R}^d)$" and the statement "$\mathrm{leb}(X(\mathbf{R}_+)) > 0$ with positive probability."

If ℓ exists and is almost surely in $L^2(\mathbf{R}^d)$, then we can apply (3.48) to deduce that ℓ is a random probability density on the closure of the range of $\mathbf{X}$. In particular, the closure of $\mathbf{X}$ – and hence $\mathbf{X}$ itself – must have positive Lebesgue measure almost surely.

Conversely, suppose the Lebesgue measure of $X(\mathbf{R}_+)$ is positive with positive probability. Equivalently, that $\kappa \in L^1(\mathbf{R}^d)$. I will follow Kahane [24], and use Fourier analysis to show that ℓ exists and is in $L^2(\mathbf{R}^d)$ almost surely.

Because $\hat{Q}(\xi) = \int_0^\infty \exp\{-s + i\xi \cdot X(s)\}\, \mathrm{d}s$ for all $\xi \in \mathbf{R}^d$, we can write $\mathrm{E}(|\hat{Q}(\xi)|^2)$ as $T_1 + T_2$, where

$$\begin{aligned} T_1 &:= \iint\limits_{0<s<t<\infty} \mathrm{e}^{-s-t}\, \mathrm{E}\left(\mathrm{e}^{i\xi\cdot[X(s)-X(t)]}\right) \mathrm{d}s\, \mathrm{d}t, \\ T_2 &:= \iint\limits_{0<t<s<\infty} \mathrm{e}^{-s-t}\, \mathrm{E}\left(\mathrm{e}^{i\xi\cdot[X(s)-X(t)]}\right) \mathrm{d}s\, \mathrm{d}t. \end{aligned} \tag{3.49}$$

If $s < t$, then the distribution of $X(s) - X(t)$ is the same as that of $-X(t-s)$, and hence

$$\begin{aligned} T_1 &= \iint\limits_{0<s<t<\infty} \mathrm{e}^{-s-t}\mathrm{E}\left(\mathrm{e}^{-i\xi\cdot X(t-s)}\right) \mathrm{d}s\, \mathrm{d}t \\ &= \iint\limits_{0<s<t<\infty} \mathrm{e}^{-s-t}\, \mathrm{e}^{-(t-s)\overline{\Psi(\xi)}}\, \mathrm{d}s\, \mathrm{d}t = \frac{1}{2 + 2\overline{\Psi(\xi)}}. \end{aligned} \tag{3.50}$$

Similarly, $T_2 = \{2 + 2\Psi(\xi)\}^{-1}$, and hence,

$$\mathrm{E}\left(|\hat{Q}(\xi)|^2\right) = \mathrm{Re}\left(\frac{1}{1 + \Psi(\xi)}\right) = \kappa(\xi). \tag{3.51}$$

Because we have assumed that κ is integrable on $\mathbf{R}^d$, this proves that $\hat{Q} \in L^2(\mathbf{R}^d)$ almost surely. Plancherel's theorem ensures us that Q is almost surely absolutely continuous with respect to the Lebesgue measure, and has an almost-surely square-integrable density ℓ. □

3.5. The sum of the range of a Lévy process and a set

Let G denote a fixed Borel-measurable subset of $\mathbf{R}^d$, and $\mathbf{X}$ a Lévy process on $\mathbf{R}^d$. We wish to know when $X(\mathbf{R}_+) \oplus G$ has positive d-dimensional Lebesgue measure [with positive probability], where $A \oplus B := \{a + b : \ a \in A, b \in B\}$. There are good reasons for asking such a question. For instance, if we consider $G := \{0\}$, then this is asking for when the range of X has positive measure; and the answer is given by Theorem 3.1 in this case. Or if $\mathbf{X}$ is a "nice" Lévy process – such as the Brownian motion – then our question turns out to be equivalent to asking when $\mathrm{P}\{0 \in X(\mathbf{R}_+) \oplus G\} > 0$. If we can answer this for all Borel sets G, then by conditioning we can decide when $\mathrm{P}\{X(\mathbf{R}_+) \cap Y(\mathbf{R}_+) \neq \varnothing\} > 0$ where Y is an independent "nice" Lévy process on $\mathbf{R}^d$. That is, we can decide when the trajectories of two independent Lévy processes can intersect. There are many other applications of these ideas as well.

Theorem 3.10 (Hawkes [21]**).** *Let $\mathbf{X}$ denote a Lévy process on $\mathbf{R}^d$ and $G \subset \mathbf{R}^d$ a nonrandom Borel-measurable set. Then the Lebesgue measure of $X(\mathbf{R}_+) \oplus G$ is positive with positive probability iff there exists a compactly supported probability measure ν on G such that*

$$\int_{\mathbf{R}^d} \kappa(\xi) |\hat{\nu}(\xi)|^2 \, \mathrm{d}\xi < \infty, \tag{3.52}$$

where κ was defined in (3.2).

The method of proof implies the following quantitative improvement:

$$\begin{aligned}\sup_{b>0} \mathrm{e}^{-b} \mathrm{E}\left[\mathrm{leb}\left(X[0\,,b] \oplus G\right)\right] &\leq \left[\frac{1}{(2\pi)^d} \inf_{\nu \in \mathcal{P}(G)} \int_{\mathbf{R}^d} \kappa(\xi) |\hat{\nu}(\xi)|^2 \, \mathrm{d}\xi\right]^{-1} \\ &\leq \mathrm{E}\left[\mathrm{leb}\left(X(\mathbf{R}_+) \oplus G\right)\right],\end{aligned} \tag{3.53}$$

where $\inf \varnothing := \infty$, and $1/\infty := 0$. Clearly, Theorem 3.10 is a consequence of (3.52).

Example. Condition (3.52) is frequently a fractal and/or capacity condition on G. For instance, consider the case that $\mathbf{X}$ is an *isotropic stable process* with index $\alpha \in (0\,,2]$. That is, $\Psi(\xi) := \text{const}\cdot\|\xi\|^\alpha$; when $\alpha = 2$ this means that $\mathbf{X}$ is Brownian motion. One can easily check that (3.52) holds if and only if $\int_{\mathbf{R}^d} \|\xi\|^{-\alpha} |\hat{\nu}(\xi)|^2 \, \mathrm{d}\xi < \infty$. Thus, a little Fourier analysis [50, Theorem 5, p. 73] shows that, in the present setting, (3.52) is equivalent to the condition that $I_{d-\alpha}(\nu) < \infty$ for some $\nu \in \mathcal{P}(G)$, where $I_q(\nu)$ is the same Riesz energy that was defined earlier in (1.8). In particular, Frostman's theorem (Theorem 1.1) implies that, in this example, $X(\mathbf{R}_+) \oplus G$ can have positive Lebesgue measure if $\dim_{\mathrm{H}} G > d - \alpha$, but not if $\dim_{\mathrm{H}} G < d - \alpha$. This finding is essentially due to McKean [37]. □

The most natural proof of Theorem 3.10 requires developing too much analytic/probabilistic machinery. Instead I will prove a close variant which has fewer requirements [though it *does* assume a good knowledge of abstract harmonic analysis at the level of Loomis [32], Pontryagin [42], or Rudin [47].]

Let Γ denote a separable *compact* metric abelian group, metrizable by a distance d which is compatible with the group structure of Γ. As is customary for abelian groups, we denote the identity of Γ by "0," the inverse of $g \in \Gamma$ by $-g$, and group multiplication by "+." We denote the Haar measure on Γ by m, using the standard normalization, $m(\Gamma) = 1$ [32, 42, 47].

Let $\mathbf{Y} := \{Y(t)\}_{t \geq 0}$ be a Lévy process with values on Γ. That is:

1. $Y(0) = 0$ almost surely;
2. $Y(t+s) - Y(s)$ is independent of $\{Y(u)\}_{0 \leq u \leq s}$ for all $s, t \geq 0$;
3. The distribution of $Y(t+s) - Y(s)$ does not depend on s, for all $s, t \geq 0$; and
4. $t \mapsto Y(t)$ is continuous in probability; i.e., $\lim_{s \to t} \mathrm{P}\{d(X(s), X(t)) > \epsilon\} = 0$ for all $t \geq 0$ and $\epsilon > 0$.

Define Γ^* to be the dual group to Γ; every *character* $\xi \in \Gamma^*$ can be identified with a one-to-one continuous mapping from Γ onto the unit disc in $\mathbf{C}$ such that $\xi(x+y) = \xi(x)\xi(y)$. It is well known that because Γ is compact, Γ^* is discrete/countable. The distribution of the entire process $\mathbf{Y}$ is determined uniquely by a function $\psi : \Gamma^* \to \mathbf{C}$ that satisfies the following:

$$\mathrm{E}\left[\xi(Y(t))\right] = \exp(-t\psi(\xi)) \qquad \text{for all } t \geq 0 \text{ and } \xi \in \Gamma^*. \tag{3.54}$$

We call ψ the *characteristic exponent of* $\mathbf{Y}$.

For all intents and purposes, you might wish to consider only the case that Γ is the torus $(0, 2\pi]^d$, in which case $\Gamma^* := \mathbf{Z}^d$ and $\xi(x) = \exp(i\xi \cdot x)$ for all $x \in \Gamma$ and $\xi \in \Gamma^*$. Then we have the following variant of Theorem 3.10:

Theorem 3.11. *Let $G \subset \Gamma$ be a nonrandom Borel-measurable set. Then the Haar measure of $Y(\mathbf{R}_+) \oplus G$ is positive with positive probability if and only if there exists a compactly-supported probability measure ν on G such that*

$$\sum_{\xi \in \Gamma^*} K(\xi) |\hat{\nu}(\xi)|^2 < \infty, \tag{3.55}$$

where

$$K(\xi) := \mathrm{Re}\left(\frac{1}{1 + \psi(\xi)}\right) \qquad \textit{for all } \xi \in \Gamma^*. \tag{3.56}$$

In fact, I will establish the following analogue of (3.52):

$$\begin{aligned} \sup_{b>0} \mathrm{e}^{-b} \mathrm{E}\left[m(Y[0, b] \oplus G)\right] &\leq \left[\inf_{\mathcal{P}(G)} \sum_{\xi \in \Gamma^*} K(\xi) |\hat{\nu}(\xi)|^2\right]^{-1} \\ &\leq \mathrm{E}\left[m(Y(\mathbf{R}_+) \oplus G)\right]; \end{aligned} \tag{3.57}$$

which appears to be a new result with novel ideas of proof. I will not prove Theorem 3.10 here. But suffice it to say that one can deduce Theorem 3.10 from Theorem 3.11 – which I *will* prove – upon first letting Γ be the large torus $[0, 2\pi n)^d$, and then "letting $n \uparrow \infty$." There are other ways of proceeding, as well.

As first step, let us recall a classical inequality.

Lemma 3.12 (Paley–Zygmund [40]**).** *If $X \in L^2(\mathrm{P})$ is nonzero with positive probability, then for all $\lambda \in [0\,,1]$,*

$$\mathrm{P}\{X \geq \lambda \mathrm{E}X\} \geq \frac{(1-\lambda)^2(\mathrm{E}X)^2}{\mathrm{E}(X^2)}. \tag{3.58}$$

Proof. If $A := \{X \geq \lambda\mathrm{E}X\}$, then

$$\mathrm{E}X = \mathrm{E}(X\mathbf{1}_A) + \mathrm{E}(X\mathbf{1}_{A^c}) \leq \sqrt{\mathrm{E}(X^2)\cdot \mathrm{P}(A)} + \lambda\mathrm{E}X, \tag{3.59}$$

by the Cauchy–Schwarz inequality. Solve for $\mathrm{P}(A)$ to finish. □

Now we can proceed with the bulk of the argument.

Proof of Theorem 3.11. It suffices to prove this theorem in the case that G is closed. We assume this condition on G henceforth.

Let P_a denote the distribution of the process $a + \mathbf{X}$ for all $a \in \Gamma$, and let E_a denote the corresponding expectation operator. We will be working with the probability measure $\mathrm{P}_m := \int_\Gamma m(\mathrm{d}a)\ \mathrm{P}_a$ and its expectation operator $\mathrm{E}_m := \int_\Gamma m(\mathrm{d}a)\ \mathrm{E}_a$.

Suppose h is a probability density on Γ, and consider

$$J(h) := \int_0^\infty h(-X(s))\mathrm{e}^{-s}\,\mathrm{d}s. \tag{3.60}$$

Since $\mathrm{E}_a[J(h)] = \int_0^\infty \mathrm{E}[h(a-X(s))]\mathrm{e}^{-s}\,\mathrm{d}s$, we integrate $[m(\mathrm{d}a)]$ to find that

$$\mathrm{E}_m[J(h)] = \int h(x)\,m(\mathrm{d}x)\cdot\int_0^\infty \mathrm{e}^{-s}\,\mathrm{d}s = 1. \tag{3.61}$$

Similarly, we can compute directly to find that

$$\mathrm{E}_m\left(|J(h)|^2\right) = 2\iint\limits_{0<s<t<\infty} \mathrm{e}^{-s-t}\,\mathrm{d}s\,\mathrm{d}t\int_\Gamma m(\mathrm{d}b)\ h(b)\mathrm{E}\left[h(b-X(t-s))\right]. \tag{3.62}$$

Since the distribution of $X(t-s)$ is μ_{t-s}, it follows that $\mathrm{E}[h(b-X(t-s))] = (\mu_{t-s}*h)(b)$, where "$*$" denotes convolution on the group algebra. After an appeal or two to the Tonnelli theorem we find that

$$\mathrm{E}_m\left(|J(h)|^2\right) = \int_\Gamma h(b)(U*h)(b)\,m(\mathrm{d}b), \tag{3.63}$$

where U is the renewal measure from (3.16). If, in addition, $h \in L^2(\Gamma)$, then $U*h \in L^2(\Gamma)$ also, and hence by Plancherel's theorem,

$$\mathrm{E}_m\left(|J(h)|^2\right) = \sum_{\xi\in\Gamma^*} \mathrm{Re}\,\hat{U}(\xi)\,|\hat{h}(\xi)|^2. \tag{3.64}$$

Because $\hat{\mu}_t(\xi) = \exp(-t\psi(\xi))$, it follows that $\mathrm{Re}\,\hat{U}(\xi) = K(\xi)$, whence

$$\mathrm{E}_m\left(|J(h)|^2\right) = \sum_{\xi\in\Gamma^*} K(\xi)\,|\hat{h}(\xi)|^2. \tag{3.65}$$

This, (3.61) and Lemma 3.12 together imply that

$$\mathrm{P}_m\{J(h)>0\} \geq \frac{1}{\sum_{\xi\in\Gamma^*} K(\xi)|\hat{h}(\xi)|^2}. \tag{3.66}$$

Now consider a function h of the form $h(x) := (\nu * \phi_\epsilon)(x)$, where: (i) $\nu \in \mathcal{P}(G)$; and (ii) $\{\phi_\epsilon\}_{\epsilon>0}$ is a continuous [compactly-supported] approximation to the identity. If $J(\nu*\phi_\epsilon) > 0$, then certainly $-X(s) \in G^\epsilon$ for some $s>0$, where G^ϵ denotes the ϵ-enlargement of G. Since $|\hat{h}(\xi)| \leq |\hat{\nu}(\xi)|$, we obtain the following after we let $\epsilon \downarrow 0$:

$$\mathrm{P}_m\left\{\overline{-Y(\mathbf{R}_+)} \cap G \neq \varnothing\right\} \geq \left[\inf_{\nu\in\mathcal{P}(G)} \sum_{\xi\in\Gamma^*} K(\xi)|\hat{\nu}(\xi)|^2\right]^{-1}. \tag{3.67}$$

This proves the first inequality in (3.53), since we can let $\epsilon \downarrow 0$ in the following:

$$\begin{aligned}\mathrm{P}_m\left\{-Y(\mathbf{R}_+)\cap G^\epsilon \neq \varnothing\right\} &= \int_\Gamma \mathrm{P}\left\{a - Y(s) \in G^\epsilon \text{ for some } s>0\right\} m(\mathrm{d}a)\\ &= \mathrm{E}\left[m\left(Y(\mathbf{R}_+)\oplus G^\epsilon\right)\right].\end{aligned} \tag{3.68}$$

Now we strive to establish the second inequality in (3.53). Without loss of generality, we may assume that $\mathrm{E}[m(Y[0\,,b]\oplus G)] > 0$; for there is nothing left to prove otherwise.

In order to obtain the converse we need some jargon from stochastic analysis. Let $\mathcal{F} := \{\mathcal{F}_t\}_{t\geq 0}$ denote the filtration generated by the process $\mathbf{Y}$; we can and will assume, without any loss in generality, that $\mathcal{F}$ satisfies the "usual conditions" [10], so that in particular Doob's optional stopping theorem applies.

Let T be the first hitting time of $-G$. That is,

$$T := \inf\left\{s>0:\ -Y(s)\in G\right\}, \tag{3.69}$$

where $\inf\varnothing := \infty$, as before. Then T is a stopping time with respect to $\mathcal{F}$. For all density functions $h:\Gamma\to\mathbf{R}_+$,

$$\begin{aligned}\mathrm{E}_m\left(J(h)\,\middle|\,\mathcal{F}_T\right) &\geq \mathrm{E}_m\left(\int_T^\infty h(Y(s))\mathrm{e}^{-s}\,\mathrm{d}s\,\middle|\,\mathcal{F}_T\right)\cdot\mathbf{1}_{\{T<\infty\}}\\ &= \mathrm{e}^{-T}\mathrm{E}_m\left(\int_0^\infty h(Y(s+T))\mathrm{e}^{-s}\,\mathrm{d}s\,\middle|\,\mathcal{F}_T\right)\cdot\mathbf{1}_{\{T<b\}}\end{aligned} \tag{3.70}$$

We apply the strong Markov property at time T to find that

$$\mathrm{E}_m\left[h(Y(s+T))\,\middle|\,\mathcal{F}_T\right] = (\mu_s * h)(Y(T)) \qquad \text{almost surely on } \{T<\infty\}, \tag{3.71}$$

where μ_s denotes the distribution of $Y(s)$ now. Consequently,

$$\mathrm{E}_m\left(J(h)\,\middle|\,\mathcal{F}_T\right) \geq \mathrm{e}^{-b}(U*h)(Y(T))\cdot\mathbf{1}_{\{T<b\}}, \tag{3.72}$$

where U is the renewal measure, defined by (3.16). Since $\mathrm{E}_m(J(h))=1$, an appeal to Doob's optional stopping theorem yields the following:

$$1 \geq \mathrm{e}^{-b}\mathrm{E}\left[(U*h)(Y(T))\,\middle|\,T<b\right]\cdot\mathrm{P}_m\{T<b\}. \tag{3.73}$$

Let $\rho(\bullet) := \mathrm{P}_m\{Y(T) \in \bullet \,|\, T < b\}$. Thus, we have

$$\mathrm{E}\left[m(Y[0\,,b] \oplus G)\right] = \mathrm{P}_m\{T < b\} \leq \frac{\mathrm{e}^b}{\int (U * h)\,\mathrm{d}\rho}. \tag{3.74}$$

[The identity follows as in (3.68).] Since U is a probability measure on Γ, if $h \in L^2(\Gamma)$, then we can apply Plancherel's theorem to find that $\int_\Gamma (U * h)\,\mathrm{d}\rho = \sum_{\xi \in \Gamma^*} \hat{U}(\xi)\hat{h}(\xi)\overline{\hat{\rho}(\xi)}$, and hence

$$\mathrm{E}\left[m(Y[0\,,b] \oplus G)\right] \leq \frac{\mathrm{e}^b}{\mathrm{Re}\sum_{\xi \in \Gamma^*} \hat{U}(\xi)\hat{h}(\xi)\overline{\hat{\rho}(\xi)}}. \tag{3.75}$$

Since Γ is compact, the preceding holds for all continuous functions h, for example. Now consider $h := \breve{\rho} * \phi_\epsilon * \breve{\phi}_\epsilon$, where: (i) $\breve{f}(x) := f(-x)$ for all functions $f : \Gamma \to \mathbf{R}$ and $x \in \Gamma$; (ii) $\breve{\rho}(A) := \rho(-A)$ for all Borel sets $A \subset \Gamma$; and (iii) $\{\phi_\epsilon\}_{\epsilon > 0}$ is an approximation to the identity comprised of all continuous functions. Thus, we obtain

$$\begin{aligned}\mathrm{E}\left[m(Y[0\,,b] \oplus G)\right] &\leq \frac{\mathrm{e}^b}{\sum_{\xi \in \Gamma^*} \mathrm{Re}\,\hat{U}(\xi) \cdot |\hat{\phi}_\epsilon(\xi)|^2 \cdot |\hat{\rho}(\xi)|^2} \\ &= \frac{\mathrm{e}^b}{\sum_{\xi \in \Gamma^*} K(\xi) \cdot |\hat{\phi}_\epsilon(\xi)|^2 \cdot |\hat{\rho}(\xi)|^2}.\end{aligned} \tag{3.76}$$

The second inequality in (3.53) follows from the preceding, and Fatou's lemma, upon letting $\epsilon \downarrow 0$. □

4. Linear stochastic PDEs

Let that μ_s denotes the distribution of a Lévy process $\mathbf{X}$ on $\mathbf{R}^d$. We have noted already that $\mu_{s+t} = \mu_s * \mu_t$, and therefore we can view $\{\mu_t\}_{t \geq 0}$ as a convolution semigroup of linear operators acting on $L^2(\mathbf{R}^d)$.

Define

$$\mathrm{Dom}[\mathcal{L}] := \left\{ f \in L^1_{\mathrm{loc}}(\mathbf{R}^d) : \int_{\mathbf{R}^d} |\hat{f}(\xi)|^2 \left(1 + |\Psi(\xi)|^2\right)\,\mathrm{d}\xi < \infty \right\}. \tag{4.1}$$

If $f \in \mathrm{Dom}[\mathcal{L}]$, then the dominated convergence theorem shows that for all $g \in L^2(\mathbf{R}^d)$ the following limit exists, and the ensuing computation is valid:

$$\begin{aligned}\int_{\mathbf{R}^d} g(x)(\mathcal{L}f)(x)\,\mathrm{d}x &:= \lim_{s \downarrow 0} \int_{\mathbf{R}^d} g(x) \left[\frac{(\mu_s * f)(x) - f(x)}{s}\right] \mathrm{d}x \\ &= \lim_{s \downarrow 0} \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \overline{\hat{g}(\xi)}\,\hat{f}(\xi) \left[\frac{\mathrm{e}^{-s\Psi(\xi)} - 1}{s}\right] \mathrm{d}\xi = -\int_{\mathbf{R}^d} \overline{\hat{g}(\xi)}\,\hat{f}(\xi)\,\Psi(\xi)\,\mathrm{d}\xi.\end{aligned} \tag{4.2}$$

Owing to duality, this defines a linear operator $\mathcal{L}$, mapping $\mathrm{Dom}[\mathcal{L}]$ in to $L^2(\mathbf{R}^d)$, such that: (1) $\mathcal{L}$ is the Hille–Yosida generator of $\{\mu_s\}_{s \geq 0}$ in the L^2-sense; and (2) the Fourier multiplier of $\mathcal{L}$ is $-\Psi$.

Example. I mention a few examples:

1. If $\Psi(\xi) = c\|\xi\|^\alpha$ for some $c > 0$ and $\alpha \in (0\,,2]$, then $\mathbf{X}$ is an isotropic stable process on $\mathbf{R}^d$ with index α – Brownian motion if $\alpha = 2$ – and $\mathcal{L} = c\Delta^{\alpha/2}$ is c times the fractional Laplacian of order $\alpha/2$.
2. When $d = 1$ and $\mathbf{X}$ is a Poisson process on $\mathbf{Z}_+ := \{0\,,1\,,2\,,\ldots\}$ with rate $\lambda > 0$, then we have $\Psi(\xi) = \lambda(1 - \mathrm{e}^{i\xi})$, and $(\mathcal{L}f)(x) = \lambda\{f(x) - f(x-1)\}\mathbf{1}_{\mathbf{Z}_+}(x)$ is λ times the discrete gradient on $\mathbf{Z}_+$.
3. When $d = 1$ and $\mathbf{X}$ is a compensated Poisson process on $\mathbf{R}$ with rate λ, $\Psi(\xi) = \lambda(1 + i\xi - \mathrm{e}^{i\xi})$, and $(\mathcal{L}f)(x) = \lambda\{f(x) - f(x-1) - f'(x)\}$ for all $x \in \mathbf{R}$.

For more information see Fukushima et al. [16] and Jacob [23]. □

The object of interest here is the so-called *stochastic heat equation,*

$$\frac{\partial u(t\,,x)}{\partial t} = (\mathcal{L}u)(t\,,x) + \dot{W}(t\,,x), \tag{4.3}$$

where $\mathcal{L}$ is the L^2-generator of a Lévy process $\mathbf{X}$ on $\mathbf{R}^d$, and $\dot{W}$ denotes white noise. That is,

$$\dot{W}(t\,,x) := \frac{\partial^{d+1}W(t\,,x)}{\partial t \partial x_1 \cdots \partial x_d}, \tag{4.4}$$

in the sense of generalized random fields, where $\mathbf{W} := \{W(t\,,x)\}_{(t,x)\in\mathbf{R}^{d+1}}$ is Brownian sheet with $d+1$ parameters. That is, $\mathbf{W}$ is a continuous centered Gaussian random field with the following covariance: For all $(t\,,x), (s\,,y) \in \mathbf{R}^{d+1}$,

$$\mathrm{Cov}\left(W(t\,,x)\ ,\ W(s\,,y)\right) = \min(s\,,t) \cdot \prod_{j=1}^{d} \min(x_j\,,y_j). \tag{4.5}$$

There are many ways to make rigorous sense of the stochastic heat equation (4.3). Here is a quick, though perhaps not the most informative, way: Let $\phi : \mathbf{R}^d \to \mathbf{R}$ be a smooth compactly-supported function from $\mathbf{R}^d$ to $\mathbf{R}$. We can multiply both sides of (4.3), purely formally, by $\phi(x)$ and integrate [dx] to arrive at the "equation,"

$$\frac{\partial u(t\,,\phi)}{\partial t} = (\mathcal{L}u)(t\,,\phi) + \dot{W}(t\,,\phi), \tag{4.6}$$

where $g(t\,,\phi) := \int_{\mathbf{R}^d} g(t\,,x)\phi(x)\,\mathrm{d}x$, whenever this makes sense, and $\dot{W}(t\,,\phi)\,\mathrm{d}t = \mathrm{d}X_t$ – as generalized Gaussian random fields – where X is a Brownian motion with covariance function

$$\mathrm{E}\left(X_s X_t\right) = \|\phi\|^2_{L^2(\mathbf{R}^d)} \cdot \min(s\,,t). \tag{4.7}$$

It is now possible to convince oneself that (4.6) ought to be interpreted as an infinite family of correlated stochastic differential equations, one for each nice ϕ. If the ensuing solution $u(t\,,\phi)$ can indeed be written as $\int_{\mathbf{R}^d} u(t\,,x)\phi(x)\,\mathrm{d}x$, then $\{u(t\,,x)\}_{t\geq 0, x\in\mathbf{R}^d}$ is a "random-field solution." References [6, 8, 25, 29, 30, 43, 46, 52] contain ways of interpreting (4.3) and many other stochastic PDEs.

We will interpret (4.3) Fourier-analytically, and prove the following:

Theorem 4.1 (Dalang [9]**, Khoshnevisan, Foondun, and Nualart** [14]**).** *Let* $\mathbf{X}'$ *denote an independent copy of* $\mathbf{X}$*, and consider the Lévy process* $\mathbf{Y} := \{Y(t)\}_{t\geq 0}$ *where* $Y(t) := X(t) - X'(t)$ *for all* $t \geq 0$*. Then,* (4.3) *has a random-field solution* $\{u(t\,,x)\}_{t\geq 0, x\in\mathbf{R}^d}$ *iff the range of* $\mathbf{Y}$ *has square-integrable local times* $\{\ell(x)\}_{x\in\mathbf{R}^d}$*.*

In particular, (4.3) never has random-field solutions when $d \geq 2$ (Remark 3.2).

We prove Theorem 4.1 after we discuss the meaning of (4.3) in detail; we shall see that the proof is based on simple ideas. Let us also mention the following deeper result whose proof is too difficult to be included here.

Theorem 4.2 (Khoshnevisan, Foondun, and Nualart [14]**).** *Suppose* $d = 1$ *and* (4.3) *has a random-field solution* $\{u(t\,,x)\}_{t\geq 0, x\in\mathbf{R}}$*. Then the following are equivalent:*

- $x \mapsto u(t\,,x)$ *is continuous with positive probability for some* $t > 0$*;*
- $x \mapsto u(t\,,x)$ *is continuous almost surely for some* $t > 0$*;*
- $x \mapsto u(t\,,x)$ *is continuous almost surely for all* $t > 0$*;*
- $x \mapsto \ell(x)$ *is almost surely continuous.*

The preceding continues to hold if "continuous" is replaced by "Hölder continuous" everywhere, and the critical index of Hölder continuity of $x \mapsto u(t\,,x)$ *is the same as that of* $x \mapsto \ell(x)$*.*

There is a large literature that describes the continuity of local times of Lévy processes. This literature culminates in the work of Barlow [1]. In order to describe that work let us consider the following function:

$$\delta(x\,,y) := \int_{-\infty}^{\infty} \frac{1 - \cos((x-y)\xi)}{1 + 2\mathrm{Re}\,\Psi(\xi)}\,\mathrm{d}\xi. \tag{4.8}$$

We can compare Theorem 4.1 with Theorem 3.9 to see that (4.3) has random-field solutions if and only if $\delta(x\,,x) < \infty$ for all x. It follows easily from this that δ is a [pseudo-] metric on $\mathbf{R}$. Define $N_\delta(\epsilon)$ to be the smallest number of δ-balls of radius $\epsilon > 0$ that are needed to cover $[0\,,1]$; N_δ is the so-called *Kolmogorov metric entropy* of $[0\,,1]$.[6] Then we have

Theorem 4.3 (Barlow [1]**).** *Suppose the symmetrized Lévy process* $\mathbf{Y}$ *has square-integrable local times* $\{\ell(x)\}_{x\in\mathbf{R}}$*. Then* $x \mapsto \ell(x)$ *is continuous almost surely if and only if it is continuous with positive probability. And the following is a necessary and sufficient condition for that continuity:*

$$\int_{0^+} \left(\log N_\delta(\epsilon)\right)^{1/2} \mathrm{d}\epsilon < \infty. \tag{4.9}$$

Barlow's theorem provides an analytic condition for the continuity of $x \mapsto \ell(x)$. According to Theorem 4.2, the same condition is necessary and sufficient for the continuity of the solution to (4.3) in its space variable x. In light of footnote

[6] We can replace $([0\,,1]\,,\mathrm{d})$ by a general compact metric space (K,d). And N_{d} is defined in the same way. If we apply it with a compact set $K \subset \mathbf{R}^d$ where d := the usual Euclidean metric instead of the preceding one, then $\limsup_{\epsilon\downarrow 0} \log N_\delta(\epsilon)/\log(1/\epsilon) = \overline{\dim}_{\mathrm{M}} K$.

6, eq. (4.9) is a non-Euclidean fractal-like condition, on the pseudo-metric space $([0\,,1]\,,\delta)$, that is necessary and sufficient for the continuity of solutions to the stochastic heat equation. [In fact, we shall see that there is basically only one good solution to (4.3) if there are any.] We close this article by making sense of (4.3) and subsequently proving Theorem 4.1.

4.1. White noise and Wiener integrals

Let $\mathcal{T}_d := [0\,,\infty) \times \mathbf{R}^d$ denote "space-time." We write a typical element of $\mathcal{T}_d$ as $(t\,,x)$ where $t \geq 0$ and $x \in \mathbf{R}^d$.

Consider the standard [complex] Hilbert space $H := L^2(\mathcal{T}_d)$, endowed with the usual inner product $\langle g\,,h\rangle_H := \int_0^\infty \mathrm{d}t \int_{\mathbf{R}^d} \mathrm{d}x\; g(t\,,x)\overline{h(t\,,x)}$ and norm $\|h\|_H := \langle h\,,h\rangle_H^{1/2}$ for all $g,h \in H$. The *isonormal process* $\mathbf{W} := \{W(h)\}_{h\in H}$ is a [complex-valued] mean-zero gaussian process whose covariance function is described by

$$\mathrm{Cov}(W(h)\,,W(g)) := \langle h\,,g\rangle_H. \tag{4.10}$$

It is not difficult to prove that for all $a,b \in \mathbf{C}$ and $h,g \in H$ the following holds almost surely: $W(ah+bg) = aW(h)+bW(g)$. But the null set depends on a, b, h, and g. In particular, if $\mathrm{Im}f \equiv 0$ then $W(f)$ is real valued, and the restriction of $\mathbf{W}$ to such functions is a *real-valued isonormal process.*

For all nonrandom Borel sets $A \subset \mathcal{T}_d$ with finite Lebesgue measure define

$$\dot{W}(A) := W(\mathbf{1}_A). \tag{4.11}$$

The resulting set-indexed stochastic process $\dot{\mathbf{W}}$ is called *white noise* on $\mathcal{T}_d$. Thus, we can think of $h \mapsto W(h)$ as an integral against white noise, and write

$$W(h) = \int_0^\infty \int_{\mathbf{R}^d} h(t\,,x)\,\dot{W}(\mathrm{d}t\,\mathrm{d}x) \qquad \text{for all } h \in H. \tag{4.12}$$

This is called the *Wiener integral* of h. We will stop writing the dot in $\dot{W}$ from here on, as there is no ambiguity in omitting that dot. Thus, from now on, we write

$$W(h) := \int_0^\infty \int_{\mathbf{R}^d} h(t\,,x)\,W(\mathrm{d}t\,\mathrm{d}x) \quad \text{for all } h \in H. \tag{4.13}$$

4.2. The Fourier transform of white noise

If $h \in H$, then so is $\hat{h}$, where $\hat{h}(t\,,\xi) := \int_{\mathbf{R}^d} \mathrm{e}^{i\xi\cdot x}h(t\,,x)\,\mathrm{d}x$ for all $\xi \in \mathbf{R}^d$; $\hat{h}$ is well defined for almost all $t \geq 0$. And we can define the *Fourier transform* $\hat{W}$ *of white noise* as

$$\hat{W}(h) := W(\hat{h}). \tag{4.14}$$

Thus, $\{\hat{W}(h)\}_{h\in H}$ is a mean-zero [complex-valued] gaussian random field whose covariance function is defined by

$$\mathrm{Cov}\left(\hat{W}(f)\,,\hat{W}(g)\right) = \langle \hat{f},\hat{g}\rangle_H = (2\pi)^d\langle f,g\rangle_H; \tag{4.15}$$

the last assertion follows from Plancherel's theorem when $f(t\,,x)$ is of the form $T(t)X(x)$, and in general by density.

4.3. A return to the linear stochastic heat equation

Before we make rigorous sense of (4.3), let us recall a few facts about the linear heat equation of classical PDEs.

Suppose $w(t\,,x)$ defines a "nice" function, and consider the heat equation

$$\frac{\partial u(t\,,x)}{\partial t} = (\mathcal{L}u)(t\,,x) + w(t\,,x). \tag{4.16}$$

We can make sense of (4.16) by taking the Fourier transform [in x] throughout. This yields $\partial \hat{u}(t\,,\xi)/\partial t = -\Psi(\xi)\hat{u}(t\,,\xi) + \hat{w}(t\,,\xi)$. Thus, a reasonable solution u to (4.16) ought to satisfy

$$\hat{u}(t\,,\xi) = \int_0^t \mathrm{e}^{-(t-s)\Psi(\xi)}\,\hat{w}(s\,,\xi)\,\mathrm{d}s. \tag{4.17}$$

In particular, we might expect that if $\phi : \mathbf{R}^d \to \mathbf{R}$ is also "nice", then

$$\int_{\mathbf{R}^d} \hat{u}(t\,,\xi)\,\overline{\hat{\phi}(\xi)}\,\mathrm{d}\xi = \int_0^t \mathrm{d}s \int_{\mathbf{R}^d} \mathrm{d}\xi\ \overline{\hat{\phi}(\xi)}\,\mathrm{e}^{-(t-s)\Psi(\xi)}\,\hat{w}(s\,,\xi). \tag{4.18}$$

An informal appeal to the Plancherel theorem might then suggest that

$$u(t\,,\phi) = \frac{1}{(2\pi)^d}\int_0^t \mathrm{d}s \int_{\mathbf{R}^d} \mathrm{d}\xi\ \overline{\hat{\phi}(\xi)}\,\mathrm{e}^{-(t-s)\Psi(\xi)}\,\hat{w}(s\,,\xi), \tag{4.19}$$

where $u(t\,,\phi) := \int_{\mathbf{R}^d} u(t\,,x)\phi(x)\,\mathrm{d}x$. A remarkable feature of this heuristic computation is that it produces the usual notion of weak solutions of (4.16) rigorously, for instance when ϕ is in the Wiener algebra $L^1(\mathbf{R}^d) \cap \mathcal{F}^{-1}(L^1(\mathbf{R}^d))$, where $\mathcal{F}$ denotes the Fourier transform.

Let $\mathfrak{D}(\Psi)$ denote the class of Schwartz distributions ϕ on $\mathbf{R}^d$ whose Fourier transform $\hat{\phi}$ is a function and

$$\int_0^t \mathrm{d}s \int_{\mathbf{R}^d} \mathrm{d}\xi\ \left|\hat{\phi}(\xi)\mathrm{e}^{-s\Psi(\xi)}\right|^2 < \infty; \tag{4.20}$$

see Dalang [9]. Then we say that $\mathfrak{D}(\Psi)$ is the *class of natural testing distributions* for (4.3), and the *weak solution* $\{u(t\,,\phi)\}_{\phi\in\mathfrak{D}(\Psi)}$ to (4.3) is the random field [i.e., stochastic process] defined by the resulting Wiener integrals

$$u(t\,,\phi) := \frac{1}{(2\pi)^d}\int_0^t \int_{\mathbf{R}^d} \overline{\hat{\phi}(\xi)}\,\mathrm{e}^{-(t-s)\Psi(\xi)}\,\hat{W}(\mathrm{d}s\,\mathrm{d}\xi). \tag{4.21}$$

This is a well-defined random field, indexed by $t \geq 0$ and $\phi \in \mathfrak{D}(\Psi)$, because

$$\mathrm{E}\left(|u(t\,,\phi)|^2\right) = \frac{1}{(2\pi)^d}\int_0^t \mathrm{d}s \int_{\mathbf{R}^d} \mathrm{d}\xi\ \left|\hat{\phi}(\xi)\,\mathrm{e}^{-(t-s)\Psi(\xi)}\right|^2 < \infty. \tag{4.22}$$

Moreover, one can show that our "weak solution" $u(t\,,\phi)$ agrees almost surely with the much-better known "weak solution" of Walsh [52] for all $\phi \in L^2(\mathbf{R}^d)$. [We will not dwell on this connection here.]

Definition 4.4. We say that (4.3) has a *random-field solution* $\{u(t\,,x)\}_{t\geq 0, x\in\mathbf{R}^d}$ if and only if $\delta_x \in \mathfrak{D}(\Psi)$ for one, and hence, all $x \in \mathbf{R}^d$. In that case, we identify $u(t\,,x)$ with $u(t\,,\delta_x)$ for each $t \geq 0$ and $x \in \mathbf{R}^d$.

This is consistent with its analogue in PDEs. In fact, if $u(t\,,x)$ exists and is sufficiently regular, then $\int_{\mathbf{R}^d} u(t\,,x)\phi(x)\,\mathrm{d}x$ defines a version of $u(t\,,\phi)$.

Define

$$\mathcal{E}(\phi\,,\psi) := \frac{1}{(2\pi)^d}\int_{\mathbf{R}^d} \frac{\hat{\psi}(\xi)\,\overline{\hat{\phi}(\xi)}}{1+2\mathrm{Re}\,\Psi(\xi)}\,\mathrm{d}\xi, \tag{4.23}$$

for all Schwartz distributions ϕ and ψ whose Fourier transform is a function and $\mathcal{E}(\phi\,,\phi) + \mathcal{E}(\psi\,,\psi) < \infty$. Because the real part of Ψ is nonnegative, $\mathcal{E}(\phi\,,\phi) < \infty$ for all $\phi \in L^2(\mathbf{R}^d)$.

Lemma 4.5. *For all $t \geq 0$ and $\phi \in L^2(\mathbf{R}^d)$,*

$$\left(1-\mathrm{e}^{-t}\right)\,\mathcal{E}(\phi\,,\phi) \leq \mathrm{E}\left(|u(t\,,\phi)|^2\right) \leq \mathrm{e}^t\,\mathcal{E}(\phi\,,\phi). \tag{4.24}$$

In particular, $L^2(\mathbf{R}^d) \subseteq \mathfrak{D}(\Psi)$.

Proof. Note that $|\exp\{-s\Psi(\xi)\}|^2 = \exp\{-2s\mathrm{Re}\,\Psi(\xi)\}$ and $\mathrm{Re}\Psi(\xi) \geq 0$. Thus, for all $t > 0$ and $\phi \in \mathfrak{D}(\Psi)$,

$$\mathrm{E}\left(|u(t\,,\phi)|^2\right) = \frac{1}{(2\pi)^d}\int_0^t \mathrm{d}s \int_{\mathbf{R}^d} \mathrm{d}\xi\ \left|\hat{\phi}(\xi)\right|^2 \mathrm{e}^{-2s\mathrm{Re}\,\Psi(\xi)}. \tag{4.25}$$

For all $\phi \in L^2(\mathbf{R}^d)$,

$$\mathrm{E}\left(|u(t\,,\phi)|^2\right) \leq \frac{\mathrm{e}^t}{(2\pi)^d}\int_0^\infty \mathrm{d}s \int_{\mathbf{R}^d} \mathrm{d}\xi\ \left|\hat{\phi}(\xi)\right|^2 \mathrm{e}^{-s-2s\mathrm{Re}\,\Psi(\xi)} = \mathrm{e}^t\,\mathcal{E}(\phi\,,\phi). \tag{4.26}$$

In order to derive the complementary bound we note that

$$\begin{aligned}\frac{1}{1+2\mathrm{Re}\,\Psi(\xi)} &= \sum_{n=0}^\infty \int_{nt}^{(n+1)t} \mathrm{e}^{-s-2s\mathrm{Re}\,\Psi(\xi)}\,\mathrm{d}s \\ &\leq \sum_{n=0}^\infty \mathrm{e}^{-nt}\int_{nt}^{(n+1)t} \mathrm{e}^{-2s\mathrm{Re}\,\Psi(\xi)}\,\mathrm{d}s \\ &\leq \frac{1}{1-\mathrm{e}^{-t}}\cdot\int_0^t \mathrm{e}^{-2r\mathrm{Re}\,\Psi(\xi)}\,\mathrm{d}r \qquad [r := s-nt].\end{aligned} \tag{4.27}$$

The lemma follows from this and an application of (4.25). □

Let us conclude the paper by proving Theorem 4.1.

Proof of Theorem 4.1. Lemma 4.5 identifies $\mathfrak{D}(\Psi)$ with the closure of $L^2(\mathbf{R}^d)$ in the "energy norm", $\phi \mapsto \mathcal{E}(\phi\,,\phi)^{1/2}$. In particular, we have a random-field solution if and only if $\mathcal{E}(\delta_x\,,\delta_x) < \infty$. An equivalent condition is the integrability of the function $\{1+2\mathrm{Re}\,\Psi\}^{-1}$ on $\mathbf{R}^d$. Since $2\mathrm{Re}\,\Psi$ is the characteristic exponent of the symmetrized Lévy process $\mathbf{Y}$, Hawkes's theorem (Theorem 3.9) completes the proof. □

Acknowledgments

I would like to thank Professors Bandt, Mörters, and Zähle for encouraging me to write an exposition of this material. This material is based in part on a decade or two of collaborations with Professor Yimin Xiao. I thank him wholeheartedly for that, and for his suggestions and corrections to the present paper. Final thanks are owed to Dr. Michael Hinz, who has kindly provided me with a long list of corrections, together with many excellent suggestions for stylistic improvements.

References

[1] M.T. Barlow, *Necessary and sufficient conditions for the continuity of local time of Lévy processes*, Ann. Probab. **16**, No. 4 (1988) 1389–1427.

[2] J. Bertoin, *Subordinators*, in: Lecture Notes in Mathematics **1717**, 1–91, Springer-Verlag, Berlin, 1999.

[3] J. Bertoin, *Lévy Processes*, Cambridge University Press, Cambridge, 1998.

[4] S. Bochner, *Harmonic Analysis and the Theory of Probability*, University of California Press, Berkeley and Los Angeles, 1955.

[5] É. Borel, *Les probabilités dénombrables et leurs applications arithmétiques*, Supplemento di rend. circ. Mat. Palermo **27** (1909) 247–271.

[6] P.-L. Chow, *Stochastic Partial Differential Equations*, Chapman & Hall/CRC, Boca Raton, Fl., 2007.

[7] M. Csörgő and P. Révész, *Strong Approximations in Probability and Statistics*, Academic Press, N.Y., 1981.

[8] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, 1992.

[9] R.C. Dalang, *Extending the martingale measure stochastic integral with applications to spatially homogeneous s.p.d.e.'s*, Electron. J. Probab. **4** (1999) no. 6, 29 pp. (electronic).

[10] C. Dellacherie and P.-A. Meyer, *Probabilities and Potential B,* translated from the French by J. P. Wilson, North-Holland Publishing Co., Amsterdam, 1982.

[11] H.G. Eggleston, *The fractional dimension of a set defined by decimal properties*, Quart. J. Math. (1949) Oxford Ser. **20**, 31–36.

[12] K.J. Falconer, *Fractal Geometry*, second edition, Wiley, New York, 2003.

[13] K.J. Falconer, *Random fractals*, Math. Proc. Camb. Phil. Soc. **100** (1986) no. 3, 559–582.

[14] M. Foondun, D. Khoshnevisan, and E. Nualart, *A local-time correspondence for stochastic partial differential equations* (2008) preprint.

[15] O. Frostman, *Potentiel d'équilibre et capacité des ensembles avec quelques applications à la théorie des fonctions*, Meddel. Lunds. Univ. Mat. Sem. **3** (1935) 1–118.

[16] M., Fukushima, Y. Ōshima, and M. Takeda, *Dirichlet Forms and Symmetric Markov Processes*, Walter de Gruyter & Co., Berlin, 1994.

[17] D. Geman and J. Horowitz, *Occupation densities*, Ann. Probab. **8** (1980) no. 1, 1–67.

[18] G. Grimmett, *Random flows: network flows and electrical flows through random media*, in: Surveys in Combinatorics, 1985 (Glasgow, 1985), 59–95, London Math. Soc. Lecture Note Ser. **103** Cambridge University Press, Cambridge, 1985.

[19] F. Hausdorff, *Grundzüge der Mengenlehre*, Verlag von Veit & Comp., Leipzig, 1914.

[20] J. Hawkes, *Local times as stationary processes*, in: "From Local Times to Global Geometry, Control, and Physics" (Coventry, 1984/85), 111–120, Pitman Lecture Notes, Math. Ser. **150**, Longman Science Tech., Harlow, 1986.

[21] J. Hawkes, *Potential theory of Lévy processes*, Proc. London Math. Soc. **38** (1979) 335–352.

[22] J. Horowitz, *The Hausdorff dimension of the sample path of a subordinator*, Israel J. Math. **6** (1968) 176–182.

[23] N. Jacob, *Pseudo Differential Operators and Markov Processes. Vol. I*, Imperial College Press, London, 2001.

[24] J.-P. Kahane, *Some Random Series of Functions*, second edition, Cambridge University Press, Cambridge, 1985.

[25] G. Kallianpur and J. Xiong, *Stochastic Differential Equations in Infinite Dimensional Spaces*, Institute of Math. Statist. Lecture Notes – Monograph Series, Hayward, California, 1995.

[26] H. Kesten, *Hitting Probabilities of Single Points for Processes with Stationary Independent Increments*. Memoirs of the Amer. Math. Soc. **93**, Amer. Math. Soc., Providence, R.I., 1969.

[27] D. Khoshnevisan, *Escape rates for Lévy processes*, Stud. Sci. Math. Hung. **33** (1997) 177–183.

[28] D. Khoshnevisan and Y. Xiao, *Packing dimension of the range of a Lévy process*, Proc. Amer. Math. Soc. **136** (2008) no. 7, 2597–2607.

[29] P. Kotelenez, *Stochastic Ordinary and Stochastic Partial Differential Equations*, Springer, New York, 2008.

[30] N.V. Krylov, *On the foundations of the L_p-theory of stochastic partial differential equations*, in: Stochastic Partial Differential Equations and Applications – VII, 179–191, Lecture Notes in Pure & Appl. Math., Chapman & Hall/CRC, Boca Raton, Fl., 2006.

[31] P. Lévy, *Processus Stochastiques et Mouvement Brownien*, Gauthier–Villars, Paris, 1948.

[32] L.H. Loomis, *Harmonic Analysis*, Notes by Ethan Bolker, John W. Blattner, and Sylvan Wallach. 1965 MAA Cooperative Summer Seminar Math. Assoc. of America, Buffalo, N.Y., 1965.

[33] B. Maisonneuve, *Ensembles régénératifs, temps locaux et subordinateurs*, in: Lecture Notes in Math. **1527**, 111–235, Springer, Berlin, 1971.

[34] B.B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman & Company, San Francisco, 1982.

[35] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces*, Cambridge University Press, Cambridge, 1995.

[36] R.D. Mauldin and S.C. Williams, *Random recursive constructions: asymptotic geometric and topological properties*, Trans. Amer. Math. Soc. **295** (1986) no. 1, 325–346.

[37] H. McKean, Jr., *Hausdorff–Besicovitch dimension of Brownian motion paths*, Duke Math. J. **22** (1955) 229–234.

[38] S. Orey, *Polar sets for processes with stationary independent increments*, in: Markov Processes and Potential Theory (Proc. Sympos. Math. Res. Center, Madison, Wis., 1967), 117–126, Wiley, New York, 1967.

[39] Orey, S. and W.E. Pruitt, *Sample functions of the N-parameter Wiener process*, Ann. Probab. **1** (1973), 138–163 .

[40] R.E.A.C. Paley and A. Zygmund, *A note on analytic functions in the unit circle*, Proc. Camb. Phil. Soc. **28** (1932) 266–272.

[41] Y. Peres, *Probability on Trees*, in: Lecture Notes in Mathematics **1717**, 195–280, Springer-Verlag, Berlin, 1999.

[42] L.S. Pontryagin, *Topological Groups*, Translated from the second Russian edition by Arlen Brown, Gordon and Breach Science Publishers, Inc., New York, 1966.

[43] C. Prévôt and M. Röckner, *A Concise Course on Stochastic Partial Differential Equations*, Lecture Notes in Mathematics **1905**, Springer-Verlag, Berlin, Heidelberg, 2007.

[44] W.E. Pruitt, *The Hausdorff dimension of the range of a process with stationary independent increments*, J. Math. Mech. **19** (1969) 71–378.

[45] C.A. Rogers and S.J. Taylor, *Functions continuous and singular with respect to a Hausdorff measure*, Mathematika **8** (1961) 1–3.

[46] B.L. Rozovskiĭ, *Stochastic Evolution Systems*, Kluwer Academic Publishing Group, Dordrecht (translated from the original Russian by A. Yarkho, *Math. and Its Applications* (Soviet Series), 35), 1990.

[47] W. Rudin, *Fourier Analysis on Groups*, Reprint of the 1962 original, John Wiley & Sons, Inc., New York, 1990.

[48] K.-I. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge [translated from the 1990 Japanese original, revised by the author],1999.

[49] I.J. Schoenberg, *Metric spaces and positive definite functions*. Trans. Amer. Math. Soc. **44** (1938) no. 3, 522–536.

[50] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Fifth printing with corrections from the original in 1970, Princeton University Press, Princeton, N.J., 1986.

[51] S.J. Taylor, *The Hausdorff α-dimensional measure of Brownian paths in n-space*, Proc. Camb. Phil. Soc. **49** (1953) 31–39.

[52] J.B. Walsh, *An Introduction to Stochastic Partial Differential Equations*, in: École d'été de probabilités de Saint-Flour, XIV – 1984, Lecture Notes in Math. **1180**, 265–439, Springer, Berlin, 1986.

Davar Khoshnevisan
University of Utah, Department of Mathematics
155 South 1400 East JWB 233
Salt Lake City UT 84112–0090, USA
e-mail: davar@math.utah.edu

Part 4

Emergence of Fractals in Complex Systems

Progress in Probability, Vol. 61, 145–174

A Survey of Dynamical Percolation

Jeffrey E. Steif

*Dedicated to the memory of Oded Schramm,
who has been a great inspiration to me and with whom
it has been a great honor and privilege to work.*

Abstract. Percolation is one of the simplest and nicest models in probability theory/statistical mechanics which exhibits critical phenomena. Dynamical percolation is a model where a simple time dynamics is added to the (ordinary) percolation model. This dynamical model exhibits very interesting behavior. Our goal in this survey is to give an overview of the work in dynamical percolation that has been done (and some of which is in the process of being written up).

Mathematics Subject Classification (2000). 60K35.

Keywords. Percolation, exceptional times.

1. Introduction

Within the very large and beautiful world of probability theory, percolation theory has been an especially attractive subject, an area in which the major problems are easily stated but whose solutions, when they exist, often require ingenious methods. Dynamical percolation is a model where we add a simple time dynamics to the percolation model. It happened to have turned out that this new model is very rich and much can be said about it. It is hoped that this survey will provide a good guide to the literature from which point the reader can study the various papers on the subject of dynamical percolation.

1.1. The ordinary percolation model

In the standard percolation model, we take an infinite connected locally finite graph G, fix $p \in [0,1]$ and let each edge (bond) of G be, independently of all others, open with probability p and closed with probability $1-p$. Write π_p for this

Research partially supported by the Swedish Natural Science Research Council and the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine.

product measure. In percolation theory, one studies the structure of the connected components (clusters) of the random subgraph of G consisting of all sites and all open edges. The first question that can be asked is the existence of an infinite connected component (in which case we say percolation occurs). Writing $\mathcal{C}$ for this latter event, Kolmogorov's 0-1 law tells us that the probability of $\mathcal{C}$ is, for fixed G and p, either 0 or 1. Since $\pi_p(\mathcal{C})$ is nondecreasing in p, there exists a critical probability $p_c = p_c(G) \in [0,1]$ such that

$$\pi_p(\mathcal{C}) = \begin{cases} 0 & \text{for } p < p_c \\ 1 & \text{for } p > p_c. \end{cases}$$

At $p = p_c$, we can have either $\pi_p(\mathcal{C}) = 0$ or $\pi_p(\mathcal{C}) = 1$, depending on G. Site percolation is defined analogously where the vertices are retained independently with probability p and all the edges between retained vertices are retained.

For the d-dimensional cubic lattice, which is the graph classically studied, the standard reference is [16]. For the study of percolation on general graphs, see [41]. For a study of critical percolation on the hexagonal lattice, see [54].

While a tremendous amount is known about this model, we highlight some key facts. In the following, $\mathbb{Z}^d$ denotes the d-dimensional integers with the usual graph structure where points at distance 1 are neighbors.

- For $\mathbb{Z}^2$, Harris [22] established in 1960 that there is no percolation at 1/2 and Kesten [28] established in 1980 that there is percolation for $p > 1/2$. In particular, we see there is no percolation at the critical value for $\mathbb{Z}^2$.
- For $\mathbb{Z}^d$ with $d \geq 19$, Hara and Slade [21] established that there is no percolation at the critical value. (Percolation connoisseurs will object to this description and say that they, in addition, essentially established this result for $d \geq 7$.)
- Aizenman, Kesten and Newman [2] proved that on $\mathbb{Z}^d$, when there is percolation, there is then a unique infinite cluster. Later Burton and Keane [8] established this in much greater generality but more importantly found a much simpler proof of this theorem.

We end this subsection with an important open question.

- For $\mathbb{Z}^d$, for intermediate dimensions, such as $d = 3$, is there percolation at the critical value?

1.2. The dynamical percolation model

Häggström, Peres and Steif [19] initiated the study of dynamical percolation. In fact, Olle Häggström and I came up with the model inspired by a question that Paul Malliavin asked at a lecture I gave at the Mittag Leffler Institute in 1995. This model was invented independently by Itai Benjamini. In this model, with p fixed, the edges of G switch back and forth according to independent 2-state continuous time Markov chains where closed switches to open at rate p and open switches to closed at rate $1-p$. Clearly, π_p is the unique stationary distribution for this Markov process. The general question studied in dynamical percolation is whether, when we start with distribution π_p, there exist atypical times at which the percolation structure looks markedly different than that at a fixed time. In almost all cases, the term “markedly different” refers to the existence or nonexistence of an infinite connected component. Dynamical percolation on site percolation models is defined analogously.

It turns out that dynamical percolation is a very interesting model which exhibits quite new phenomena. It also connects up very much with the recent and important developments in 2-d critical percolation.

In the paper [17], Häggström gives a description of the early results in dynamical percolation and some open questions; most of these earlier results will be repeated here. We also mention the following upcoming paper [13] by Garban which will discuss Oded Schramm’s contributions to the area of noise sensitivity; this paper will have some overlap with the present paper.

1.3. Outline of paper

In Section 2, we will start with a discussion of some of the first results obtained for dynamical percolation including (1) “triviality away from criticality”, (2) the existence of graphs with “exceptional times” and (3) nonexistence of “exceptional times” for high-dimensional lattices. In Section 3, we present the fairly refined results that have been obtained for trees that do not percolate at criticality in ordinary percolation.

We present in Section 4 the basic elements of noise sensitivity and influence together with the Fourier spectrum which is the key tool for the analysis of these concepts. These concepts are both interesting in themselves and are key in the “second moment method” approach to proving the existence of exceptional times for dynamical percolation in 2 dimensions as well as to computing the Hausdorff dimension of the set of exceptional times. Section 5 briefly reviews results concerning critical exponents for ordinary critical percolation since these are crucial to the study of dynamical percolation in 2 dimensions.

In Section 6, we outline the proofs from [51] for the hexagonal lattice of the existence of exceptional times for dynamical percolation (which also yields a nonsharp lower bound on the Hausdorff dimension) and obtain a lower bound for the noise sensitivity exponent for crossing probabilities. (For the square lattice, positivity of the noise sensitivity exponent was established but not the existence of exceptional times.) This method managed to estimate the Fourier spectrum well enough to obtain the above two results. However, the estimates of the spectrum that this method led to were not sharp enough to yield the exact noise sensitivity exponent or the exact dimension of the set of exceptional times. We present the results from [14] in Section 7. These yield for the hexagonal lattice the exact dimension of the set of exceptional times for dynamical percolation and the exact noise sensitivity exponent for crossing probabilities. (For the square lattice, existence of exceptional times for dynamical percolation as well as sharp noise results in terms of the number of pivotals is also established.) These arguments are based on a completely different method to analyze the Fourier spectrum which turns out to be sharp.

Essentially all the results above dealt with graphs which do not percolate at criticality. When dealing with graphs which do percolate at criticality and asking if there are exceptional times of *non*-percolation, the structure of the problem is very different. A number of results in this case are presented in Section 8.

In Section 9, a relationship between the incipient infinite cluster (in ordinary percolation) and the structure of the infinite cluster at exceptional times in dynamical percolation is presented. Recent work on the scaling limit of dynamical percolation on the hexagonal lattice is discussed in Section 10, . Finally, in Section 11, dynamical percolation for certain interacting particle systems are discussed.

We end the introduction by mentioning that there are a number of other papers, such as [3], [24] and [23], dealing with the existence of certain "exceptional times" for various models. On an abstract level, the questions we are asking concerning exceptional behavior are equivalent to questions concerning the polarity or non-polarity of certain sets for our (Markovian) dynamical percolation model; see [12]. Sometimes, the exceptional objects one looks for are of a more general nature such as in [5] and [1].

2. Dynamical percolation: first results

This section will discuss a selected subset of the results that were obtained in [19] and [49], which were the first two papers on dynamical percolation.

2.1. Away from criticality

The first very elementary result is the following and contained in [19]. Let $\mathcal{C}_t$ be the event that there is an infinite cluster at time t and P_p denote the probability measure governing the whole process when parameter p is used and the graph is understood.

Proposition 2.1. [19] *For any graph G we have*

$$\mathrm{P}_p\big((\neg\,\mathcal{C}_t) \textit{ occurs for every } t\big) = 1 \quad \textit{if} \quad p < p_c(G)$$

$$\mathrm{P}_p(\,\mathcal{C}_t \textit{ occurs for every } t\,) = 1 \quad \textit{if} \quad p > p_c(G)\,.$$

Outline of proof. For the first part, we choose δ sufficiently small so that the set of edges which are open at *some time* during a time interval of length δ is still subcritical. Then on each "δ-interval" there are no exceptional times of percolation and we use countable additivity. The second part is proved in the same way. □

This result suggests that the only interesting parameter is the critical one and this is essentially true. However, before restricting ourselves in the continuation to the critical case, we do mention two results for the supercritical case, both from [49].

The first result is a version for dynamical percolation of the uniqueness of the infinite cluster.

Theorem 2.2. [49] *Consider dynamical percolation with parameter $p > p_c$ on the d-dimensional cubic lattice $\mathbb{Z}^d$. Then a.s., for all times t, there exists a unique infinite cluster.*

Proposition 2.3. [49] *Let Γ be any infinite tree. If $p \in (p_c(\Gamma), 1)$, then a.s., there exist infinitely many infinite clusters for all t.*

2.2. Graphs with exceptional times exist

We now stick to the critical case. The following result says that dynamical percolation can exhibit *exceptional times* which then insures that this model is interesting.

Theorem 2.4. [19] *There exists a G_1 which does not percolate at criticality but such that there exist exceptional times of percolation.*
There also exists a graph G_2 which percolates at criticality but such that there exist exceptional times of nonpercolation.

These graphs were obtained by replacing the edges of $\mathbb{Z}^2$ by larger and larger finite graphs which simulated having the original edge move back and forth quickly. These graphs have unbounded degree. Many examples with bounded degree were later discovered and will be discussed in later sections.

2.3. High-dimensional Euclidean case

Once we have the existence of such examples, it is natural to ask what happens on the standard graphs that we are familiar with. The following result answers this for the cubic lattice in high dimensions. (Recall Hara and Slade [21] proved that ordinary percolation does not percolate at criticality in this case.)

Theorem 2.5. [19] *For the integer lattice $\mathbb{Z}^d$ with $d \geq 19$, dynamical critical percolation has no exceptional times of percolation.*

The key reason for this is a highly nontrivial result due to Hara and Slade, that says that if $\theta(p)$ is the probability that the origin percolates when the parameter is p, then

$$\theta(p) = O(p - p_c)\,. \tag{2.1}$$

In fact, Theorem 2.5 was shown to hold under the assumption (2.1).

Outline of proof. We use a first moment argument together with the proof method of Proposition 2.1. We break the time interval $[0, 1]$ into n intervals each of length $1/n$. If we fix one of these intervals, the set of edges which are open at *some time* during this interval has density about $p_c + 1/n$. Hence the probability that the origin percolates with respect to these set of edges is by (2.1) at most $O(1/n)$. It follows that the expected number of these intervals where this occurs is at most $O(1)$. It can then be argued using Fatou's Lemma that a.s. there are at most finitely many exceptional times during $[0, 1]$ at which the origin percolates. To go from there to no exceptional times can either be done by using some rather abstract Markov process theory or, as was done in the paper, by hand, which was not completely trivial. □

It is known, due to Kesten and Zhang [32], that (2.1) fails for $\mathbb{Z}^2$. The question of whether there are exceptional times for critical dynamical percolation on $\mathbb{Z}^2$ was left open in [19]. (Recall there is no percolation at a fixed time in this case.) This question will be resolved in Sections 6 and 7.

3. Exceptional times of percolation: the tree case

Although Theorem 2.4 demonstrated the existence of graphs which have exceptional times where the percolation picture is different from that at a fixed time, in [19], a more detailed analysis was done for trees.

To explain this, we first need to give a little background from ordinary percolation on trees, results which are due to Lyons and contained in [39] and [40]. Lyons obtained an explicit condition when a given tree percolates at a given value of p. This formula becomes simplest when the (rooted) tree is so-called *spherically symmetric* which means that all vertices at a given level have the same number of children, although this number may depend on the given level.

Theorem 3.1. [40] (*Case of spherically symmetric trees*)
Let 0 be the root of our tree and T_n be the number of vertices at the nth level. Then the following hold.

(i) $p_c(T) = (\liminf_n T_n^{1/n})^{-1}$.
(ii) $\mathrm{P}_p(0 \leftrightarrow \infty) > 0$ *if and only if* $\sum_n \frac{1}{w_n} < \infty$ *where*
$w_n := \mathrm{E}_p[$*number of vertices in nth level connected to 0*$] = p^n T_n$.

Note that (i) essentially follows from (ii). In [19], an explicit condition was obtained for when, given a general tree and a value of p, there exists for dynamical percolation a time at which percolation occurs. Again the spherically symmetric case is simplest.

Theorem 3.2. [19] (*Case of spherically symmetric trees*)
Let w_n be as in the previous result. Then there exists a.s. a time at which percolation occurs (*equivalently with positive probability there exists a time in* $[0,1]$ *at which the root percolates*) *if and only if* $\sum_n \frac{1}{n w_n} < \infty$.

Example. If $T_n \asymp 2^n n^{1/2}$, then the critical value is 1/2, there is no percolation at criticality but there are exceptional times at which percolation occurs (which will have dimension 1/2 by our next result). (The symbol $\asymp$ means that the ratio of the two sides is bounded away from zero and infinity by constants.)

Outline of proof. For the "if" direction, we apply the second moment method to the random variables

$$Z_n := \sum_{v \in T_n} U_v$$

where U_v is the Lebesgue amount of time during $[0,1]$ that the root is connected to v and T_n now denotes the set of vertices at the nth level. (The "second moment method" means that one computes both the first and second moments of a given nonnegative random variable X and applies the Cauchy-Schwarz inequality to obtain $P(X > 0) \geq (E[X]^2)/E[X^2]$.) For the "only if" direction, we perform a slightly nontrivial *first space-time decomposition argument* as follows. If there is some v in T_n which percolates to the root at some time t during $[0, 1]$, we show that one can pick such a random pair (v', t') in such a way that $E[Z_n \mid (v', t')]$ is much larger than $E[Z_n]$. This implies that it is very unlikely that such a pair exists. □

In [18], it was shown that some of the more refined results in [40] and Theorem 3.2 above together yield that a spherically symmetric tree has an exceptional time of percolation if and only if $\int_{p_c}^1 1/\theta(p) < \infty$. (As we will see, this equivalence also turns out to be true for $\mathbb{Z}^d$ with $d = 2$ or $d \geq 19$.) It was also shown in [18] that for general trees, the "only if" holds but not the "if".

In the spherically symmetric case, the Hausdorff dimension was also determined in [19].

Theorem 3.3. [19] *Consider a spherically symmetric tree and let w_n be as in the two previous results. Then the Hausdorff dimension of the set of times at which percolation occurs is given by*

$$\sup\left\{ \alpha \in [0, 1] \ : \ \sum_{n=1}^{\infty} \frac{n^{\alpha-1}}{w_n} < \infty \right\}.$$

While we do not state the results here, we mention that [49] went a good deal further for spherically symmetric trees. What was obtained in [49] was (see Corollary 5.1 there) a "capacity condition" for which subsets of time have the property that with positive probability they contain a percolating time. This is analogous to the well-known Kakutani criterion (in terms of Newtonian capacity) of which subsets in Euclidean space intersect a Brownian motion path with positive probability.

Once we have such a capacity condition, Peres' intersection equivalence theory ([45] and [46]) leads to a criterion for when there are exceptional times at which there are at least k infinite clusters. We can in particular construct trees for which there are times at which we have (say) 6 infinite clusters but no times at which there are 7 infinite clusters. In addition, various Hausdorff dimensions of these different exceptional time sets can be computed.

On a much less formal note, my personal feeling is that the set of exceptional times in these cases, very vaguely speaking, might have a similar structure to the set of so-called "slow" points for Brownian motion. In Section 8, we will see a very different type of set of exceptional times for dynamical percolation and I believe that in this latter case, this set might behave more like the set of so-called "fast" points for Brownian motion. Here are two words explaining this vague connection (for those who know these concepts from Brownian motion). For a time point s to be a slow point, it must be the case that $|B(t + s) - B(s)|$ does not go above a certain (well-specified) level for *all* values of small t while a time s is an exceptional time of percolation if the origin percolates out to *all* distances. On the other hand, for a time point s to be a fast point, it must be the case that $|B(t+s)-B(s)|$ goes above a certain (well-specified) level for an *infinite number* of arbitrarily small t while a time s is an exceptional time of non-percolation say for a tree if we have an *infinite number* of cut-sets which are off. This type of structure looks like something which is called a limsup fractal; see [35].

We end by mentioning that in [33], Khoshnevisan extended to general trees the result of Peres and Steif determining which time sets contain percolating times. Results concerning the Hausdorff dimension of exceptional times for general trees are also obtained in [33]. Some of the techniques use methods from [34].

4. Noise sensitivity, noise stability, influence and the Fourier spectrum

The study of noise sensitivity and noise stability for Boolean functions was initiated in [4]. The concepts discussed in this section come from or are motivated by this source. See [44] for a nice survey of noise sensitivity and its applications in theoretic computer science. See also [26] for related matters.

4.1. Definition of noise sensitivity, noise stability and some examples

As indicated in the introduction, the notion of noise sensitivity is *both* an interesting concept in itself *and* what is needed to carry out the second moment arguments necessary for the results described in Sections 6 and 7.

Let ω be uniformly chosen from the n-dimensional discrete cube $\{0,1\}^n$ (which we can think of as n fair coin flips) and let ω_ϵ be ω but with each bit independently "rerandomized" with probability ϵ. "Rerandomized" means the bit (independently of everything else) rechooses whether it is 1 or 0, each with probability 1/2. (Of course ω_ϵ has the same distribution as ω).

Let $f : \{0,1\}^n \to \{\pm 1\}$ or $\{0,1\}$ be arbitrary. We are interested in the covariance between $f(\omega)$ and $f(\omega_\epsilon)$. In most cases of interest, we will have a sequence $\{f_n\}$ where $f_n : \{0,1\}^{m_n} \to \{\pm 1\}$ or $\{0,1\}$ and we are interested in the asymptotic behavior of the covariance above.

Definition 4.1. The sequence $\{f_n\}$ is **noise sensitive** if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{E}[f_n(\omega)f_n(\omega_\epsilon)] - \mathrm{E}[f_n(\omega)]^2 = 0.$$

Usually f is an indicator of an event A and this then says that the two events $\{\omega \in A\}$ and $\{\omega_\epsilon \in A\}$ are close to independent for ϵ fixed and n large. The following notion captures the opposite situation where the two events above are close to being the same event if ϵ is small, uniformly in n.

Definition 4.2. The sequence $\{f_n\}$ is **noise stable** if

$$\lim_{\epsilon\to 0} \sup_n \mathrm{P}(f_n(\omega) \neq f_n(\omega_\epsilon)) = 0.$$

It is easy to check that $\{f_n\}$ is both noise sensitive and noise stable if and only if the sequence of variances $\{\mathrm{Var}(f_n)\}$ goes to 0. Here are two easy examples where m_n is taken to be n.

Example 1. $f_n(\omega) = \omega_1$ (i.e., the first bit).

Example 2. $f_n(\omega)$ is the parity of the number of 1's in ω.

It is easy to check that Example 1 is noise stable while Example 2 is noise sensitive. We will see later why these two examples are the two extreme examples. A more interesting example is the following which, while it is not immediately obvious, turns out to be noise stable as shown in [4].

Example 3. (Majority) Let $m_n = 2n + 1$. Let $f_n(\omega)$ be 1 if there is a majority of 1's and 0 if there is a majority of 0's.

A much more interesting example is the following.

Example 4. Let f_n be the indicator function of a left to right crossing of the box $[0, n] \times [0, n]$ for critical percolation either for the ordinary lattice $\mathbb{Z}^2$ or for the hexagonal lattice. For the hexagonal lattice, the box has to be slightly modified so that it is a union of hexagons. (Note m_n is of order n^2 in this case.)

Theorem 4.3. [4] *The sequence* $\{f_n\}$ *in Example* 4 *is noise sensitive.*

In fact, it was shown that

$$\lim_{n\to\infty} \mathrm{E}[f_n(\omega)f_n(\omega_{\epsilon_n})] - \mathrm{E}[f_n(\omega)]^2 = 0 \tag{4.1}$$

even when ϵ_n goes to 0 with n provided that $\epsilon_n \geq C/\log(n)$ for a sufficiently large C. Clearly for any sequence of Boolean functions, if ϵ_n goes to 0 sufficiently fast, we have

$$\lim_{n\to\infty} \mathrm{P}(f_n(\omega) \neq f_n(\omega_{\epsilon_n})) = 0 \tag{4.2}$$

for the trivial reason that in that case $\mathrm{P}(\omega \neq \omega_{\epsilon_n}) \to 0$.

4.2. The noise sensitivity exponent, the noise stability exponent, and influence

For sequences $\{f_n\}$ which are noise sensitive, it might be hoped that (4.1) is still true when ϵ_n decays as some power of $1/n$ and this was explicitly asked in [4] for crossings in critical percolation. This suggested "critical exponent" is what we call the noise sensitivity exponent. The following definitions now seem natural.

Definition 4.4. The **noise sensitivity exponent** (SENS($\{f_n\}$)) of a sequence $\{f_n\}$ is defined to be

$$\sup\{\alpha : \text{ (4.1) holds with } \epsilon_n = (1/n)^\alpha\}.$$

The **noise stability exponent** (STAB($\{f_n\}$)) of a sequence $\{f_n\}$ is defined to be

$$\inf\{\alpha : \text{ (4.2) holds with } \epsilon_n = (1/n)^\alpha\}.$$

Remarks.

1. We will see later that $\mathrm{E}[f(\omega)f(\omega_\epsilon)] - \mathrm{E}[f(\omega)]^2$ is nonnegative and decreasing in ϵ. It easily follows that $\mathrm{P}(f_n(\omega) \neq f_n(\omega_\epsilon))$ is increasing in ϵ. From this, it is easy to see that $\mathrm{SENS}(\{f_n\}) \leq \mathrm{STAB}(\{f_n\})$ unless the variances $\mathrm{Var}(\{f_n\})$ go to 0.
2. One might hope that $\mathrm{SENS}(\{f_n\}) = \mathrm{STAB}(\{f_n\})$. First, this can fail for trivial reasons such as the f_n's for even n might have nothing to do with the f_n's for odd n. However, this may fail for more interesting reasons. Using $\mathrm{STAB}(\{A_n\})$ for $\mathrm{STAB}(\{I_{A_n}\})$ and similarly for SENS, if, for example, A_n and B_n are independent for each n, have probabilities near 1/2 and satisfy
$$\mathrm{SENS}(\{A_n\}) = \mathrm{STAB}(\{A_n\}) = a < b = \mathrm{SENS}(\{B_n\}) = \mathrm{STAB}(\{B_n\}),$$
then it is easy to check that
$$\mathrm{SENS}(\{A_n \cap B_n\}) = a < b = \mathrm{STAB}(\{A_n \cap B_n\}).$$
In such a case, for $\epsilon_n = (1/n)^c$ with $c \in (a, b)$, the correlation between $\{\omega \in A_n \cap B_n\}$ and $\{\omega_{\epsilon_n} \in A_n \cap B_n\}$ neither goes to 0 nor to being perfectly correlated. The question of under what conditions we have $\mathrm{SENS}(\{f_n\}) = \mathrm{STAB}(\{f_n\})$ turns out to be a fairly subtle one.
3. If $m_n \asymp n^\sigma$, then it is trivial to check that $\mathrm{STAB}(\{f_n\}) \leq \sigma$ since if $\epsilon_n = (1/n)^{\sigma+\delta}$ for some fixed δ, then $\mathrm{P}(\omega \neq \omega_{\epsilon_n}) \to 0$.
4. It is natural to ask for bounds on these exponents for general Boolean functions; this will be discussed at the end of subsection 4.3.

The next important notion of total influence will give us a more interesting upper bound on the noise stability exponent than that provided by comment 3 above.

Definition 4.5. Given a function $f : \{0,1\}^n \to \{\pm 1\}$ or $\{0,1\}$, let $I_i(f)$, which we call the **influence** of the ith variable on f, be the probability that all the variables other than the ith variable do not determine f. I.e., letting ω^i be ω but flipped in the ith coordinate,
$$I_i(f) := \mathrm{P}(\omega : f(\omega) \neq f(\omega^i)).$$
The **total influence**, denote by $I(f)$, is defined to be $\sum_i I_i(f)$. If $f(\omega) \neq f(\omega^i)$, we say that i is **pivotal** for f and ω and hence $I(f)$ is the expected number of pivotal bits of f.

The following I believe was "known" in the community. The argument below however I first saw given by Christophe Garban in the context of percolation.

Theorem 4.6. *Consider a sequence $f_n : \{0,1\}^{m_n} \to \{\pm 1\}$ or $\{0,1\}$ and assume that $I(f_n) = n^{\rho+o(1)}$. Then $\mathrm{STAB}(\{f_n\}) \leq \rho$.*

Proof. We need to show that if $\alpha > \rho$ and $\epsilon_n = (1/n)^\alpha$, then (4.2) holds. Let $\omega_0, \omega_1, \ldots, \omega_{k_n}$ be such that one obtains ω_{i+1} from ω_i by choosing independently a bit at random and rerandomizing it. It is immediate that $\mathrm{P}(f_n(\omega_i) \neq f_n(\omega_{i+1})) =$

$\frac{I(f_n)}{2m_n}$ from which one gets $\mathrm{P}(f_n(\omega_0) \neq f_n(\omega_{k_n})) \leq \frac{k_n n^{\rho+o(1)}}{m_n}$. If $\epsilon_n = (1/n)^\alpha$, then ω_{ϵ_n} is sort of like ω_{k_n} with $k_n = \frac{m_n}{n^\alpha}$ and so

$$\mathrm{P}(f_n(\omega) \neq f_n(\omega_{\epsilon_n})) \sim \mathrm{P}(f_n(\omega_0) \neq f_n(\omega_{k_n})) \leq \frac{m_n n^{\rho+o(1)}}{m_n n^\alpha}$$

which goes to 0 as $n \to \infty$ if $\alpha > \rho$. The "sort of like" above and the imprecise $\sim$ are trivial to make rigorous and correct using standard large deviations for binomial random variables. □

Remarks. An example where we have strict inequality is the "majority function" of Example 3. It is easy to check that for this example ρ in Theorem 4.6 is 1/2 but, by the noise stability of this example, we have that $\mathrm{STAB}(\{f_n\}) = 0$. One explanation of the failure of having a converse to Theorem 4.6 in this case is that the expected number of pivotals is not so relevant: the random number N_n of pivotals is not at all concentrated around its mean $\mathrm{E}[N_n] = I(f_n) \asymp n^{1/2}$ but rather it goes to 0 in probability.

We will see an alternative proof of Theorem 4.6 using Fourier analysis in the next subsection.

Of the very many nice results in [4], we mention the following one which almost characterizes noise sensitivity in terms of influences. As the authors mention, this result could be used to prove Theorem 4.3 but they instead use a different approach. A function f is **monotone** if $x \leq y$ (meaning $x_i \leq y_i$ for each i) implies that $f(x) \leq f(y)$. The proof of the following result uses the Fourier spectrum (see the next subsection) and a notion known as hypercontractivity.

Theorem 4.7. [4] *Consider a sequence of Boolean functions $\{f_n\}$.*

If $\lim_{n\to\infty} \sum_i I_i(f_n)^2 = 0$, then the sequence $\{f_n\}$ is noise sensitive. The converse is true if the f_n's are monotone. (*Example* 2 *shows that monotonicity is needed for the converse.*)

We end this section by going back to Example 4. It was asked in [4] whether the noise sensitivity exponent for this sequence is 3/4. The heuristic for this guess is the following. An edge (or hexagon) is pivotal if and only if there are 4 disjoint monochromatic paths alternating in color from the edge (or hexagon) to the top, right, bottom and left sides of the box. This is close to the 4-arm exponent which by [53] was proved to be behave like $n^{-5/4}$. If boundary terms do not matter much, $I(f_n)$ should then be about $n^{3/4}$. One might hope that the number N_n of pivotals is quite concentrated around its mean $I(f_n)$; in this direction, it is well known for example that $\mathrm{E}[N_n^2] = O(1)\mathrm{E}[N_n]^2$. This gives hope that Theorem 4.6 is now tight in this case and that 3/4 might be both the noise sensitivity and noise stability exponents. In Section 7, we will see that this is indeed the case, but the proof will not go through understanding pivotals but rather through Fourier analysis. This brings us to our next topic.

4.3. The Fourier spectrum

It turns out that the best and proper way to analyze the above problems is to use Fourier analysis. The set of all functions $f : \{0,1\}^n \to \mathbb{R}$ is a 2^n-dimensional vector space with orthogonal basis $\{\chi_S\}_{S\subseteq\{1,\dots,n\}}$ where

$$\chi_S(\omega^1,\dots,\omega^n) := \begin{cases} -1 & \text{if } \#\text{ of 1's in } \{\omega^i\}_{i\in S} \text{ is odd} \\ 1 & \text{if } \#\text{ of 1's in } \{\omega^i\}_{i\in S} \text{ is even.} \end{cases}$$

So $\chi_\emptyset \equiv 1$. We then can write

$$f := \sum_{S\subseteq\{1,\dots,n\}} \hat{f}(S)\chi_S.$$

(In fancy terms, the various χ_S's are the so-called characters on the group $\mathbb{Z}_2^n$ but everything below will be from first principles.) The $\hat{f}(S)$'s are called the Fourier coefficients.

The reason that $\{\chi_S\}$ is a useful basis is that they are eigenfunctions for the discrete time Markov chain which takes ω to ω_ϵ. It is an easy exercise to check that

$$\mathrm{E}[\chi_S(\omega_\epsilon)|\omega] = (1-\epsilon)^{|S|}\chi_S(\omega)$$

and that

$$\mathrm{E}[f(\omega)f(\omega_\epsilon)] = \mathrm{E}[f(\omega)]^2 + \sum_{k=1}^{n}(1-\epsilon)^k \sum_{|S|=k} \hat{f}(S)^2. \tag{4.3}$$

This formula which first (I believe) appeared in [4] makes clear the central role played by the Fourier coefficients with respect to questions involving noise. Note that we see the nonnegativity of the covariance between $f(\omega)$ and $f(\omega_\epsilon)$ and that it is decreasing in ϵ as we had claimed earlier. (I am not aware of any coupling proof of this latter fact; we can of course couple ω, ω_ϵ and $\omega_{\epsilon+\delta}$ so that ω_ϵ agrees with ω in more places than $\omega_{\epsilon+\delta}$ does but in view of Example 2 in Subsection 4.1, this does not seem to help.) Crucially, we see that the covariance between $f(\omega)$ and $f(\omega_\epsilon)$ is small when most of the "weight" of these coefficients are on the S's with larger cardinality while the covariance is largest when most of the "weight" of these coefficients are on the smaller S's.

In Subsection 4.1, Example 1 is the function $(1-\chi_{\{1\}})/2$ while Example 2 is the function $\chi_{\{1,\dots,n\}}$ from which we now see why these are extreme examples as we mentioned earlier.

We now restrict to f's which take values in $\{-1,1\}$. The advantage in doing this is that we have (due to the Pythagorean theorem or Parseval's formula)

$$\sum_{S\subseteq\{1,\dots,n\}} \hat{f}(S)^2 = 1.$$

(For those who do not like the restriction of $\{-1,1\}$ on the range of f since you are interested in indicator functions of events, you can just consider the function which is 1 on the event in question and -1 on its complement and then easily translate back the results below in terms of your events.)

Given the above, we can let $\mathcal{S}$ be a random subset of $\{1,\ldots,n\}$ given by $\mathrm{P}(\mathcal{S}=S)=\hat{f}(S)^2$. The idea of looking at this as a probability distribution on the subsets of $\{1,\ldots,n\}$ was proposed in [4] and is called the **spectral sample** or **spectral measure** of f. (We will not be careful to distinguish between $\mathcal{S}$ and its distribution.) (4.3) can now be written as (with the two expectations being on different spaces)

$$\mathrm{E}[f(\omega)f(\omega_\epsilon)]=\mathrm{E}[(1-\epsilon)^{|\mathcal{S}|}]. \tag{4.4}$$

This equation demonstrates the important fact that a sequence $\{f_n\}$ is noise sensitive if and only if the corresponding spectral measures $\{\mathcal{S}_n\}$ satisfy $|\mathcal{S}_n|\to\infty$ in distribution provided we remove the point mass at $\emptyset$ (which corresponds to subtracting the squared mean).

We now give an alternative proof of Theorem 4.6 which is taken from [14].

Proof [14]. We need only one thing which we do not prove here which comes from [25]; one can also see a proof of this in [14]. This is that

$$\mathrm{E}[|\mathcal{S}|]=I(f). \tag{4.5}$$

It clearly suffices to show that if $\alpha>\rho$ and $\epsilon_n=(1/n)^\alpha$, then

$$\lim_{n\to\infty}\mathrm{E}[f_n(\omega)f_n(\omega_{\epsilon_n})]=1.$$

(4.4), Jensen's inequality and $\mathrm{E}[|\mathcal{S}_n|]=I(f_n)$ yields

$$\begin{aligned}\mathrm{E}[f_n(\omega)f_n(\omega_{\epsilon_n})]&=\mathrm{E}[(1-\epsilon_n)^{|\mathcal{S}_n|}]\ge(1-\epsilon_n)^{\mathrm{E}[|\mathcal{S}_n|]}=(1-\epsilon_n)^{I(f_n)}\\&=(1-(1/n)^\alpha)^{n^{\rho+o(1)}}.\end{aligned}$$

This goes to 1 since $\alpha>\rho$. □

In the context of Theorem 4.6, it would be nice to know under which further conditions we could conclude the reverse inequality and even that $\mathrm{SENS}(\{f_n\})\ge\rho$. One sufficient condition is that the distributions of $|\mathcal{S}_n|$, normalized by their means and with the point mass at 0 removed is tight on $(0,\infty)$. (Tightness at ∞ follows from Markov's inequality; the key point is tightness near 0.) We make this precise in the following result. This result is proved by just abstracting an easy small part of an argument from [14].

Theorem 4.8. *Assume $f_n:\{0,1\}^{m_n}\to\{\pm1\}$ and that $I(f_n)=n^{\rho+o(1)}$. Let $\mathcal{S}_n$ be the spectral sample corresponding to f_n. Assume that for every $\gamma>0$, there is $\delta>0$ so that for all n*

$$\mathrm{P}(|\mathcal{S}_n|<\delta\mathrm{E}[|\mathcal{S}_n|],\mathcal{S}_n\ne\emptyset)<\gamma. \tag{4.6}$$

Then $\mathrm{SENS}(\{f_n\})\ge\rho$.

Proof. We need to show that if $\alpha<\rho$, (4.1) holds when $\epsilon_n=(1/n)^\alpha$. The difference in (4.1) is by (4.3) and (4.4) simply $\mathrm{E}[(1-\epsilon_n)^{|\mathcal{S}_n|}I_{\mathcal{S}_n\ne\emptyset}]$. Fix $\gamma>0$ and choose δ as in the assumption. We have

$$\mathrm{E}[(1-\epsilon_n)^{|\mathcal{S}_n|}I_{\mathcal{S}_n\ne\emptyset}]\le\mathrm{P}(|\mathcal{S}_n|<\delta\mathrm{E}[|\mathcal{S}_n|],\mathcal{S}_n\ne\emptyset)+(1-\epsilon_n)^{\delta\mathrm{E}[|\mathcal{S}_n|]}.$$

The first term is at most γ for all n by the choice of δ and the last term goes to 0 since δ is fixed, $\alpha < \rho$ and using (4.5). Since γ was arbitrary, we are done. □

It is now interesting to look again at the majority function for which $\rho = 1/2$ but $\text{SENS}(\{f_n\}) = \text{STAB}(\{f_n\}) = 0$. In this case, the spectral measures do not satisfy the necessary tightness condition above but rather these distributions normalized by their means approach the point mass at 0. This follows from the known noise stability of majority, the fact (see Theorem 1.9 in [4]) that stability in general implies tightness at ∞ for the *unnormalized* spectral measures and the fact that the expected value of the spectral size is going to ∞. See [44] for details concerning the spectral measure of majority.

The following is an exercise for the reader which relates the lower tail of the spectral measures with noise sensitivity; Theorem 4.8 already gave some relationship between these.

Exercise 1. Let f_n be an arbitrary sequence of Boolean functions with corresponding spectral samples $\mathcal{S}_n$.

(i) Show that $\mathrm{P}(0 < |\mathcal{S}_n| \le A_n) \to 0$ implies that $\mathrm{E}[(1-\epsilon_n)^{|\mathcal{S}_n|} I_{\mathcal{S}_n \neq \emptyset}] \to 0$ if $\epsilon_n A_n \to \infty$.

(ii) Show that $\mathrm{E}[(1-\epsilon_n)^{|\mathcal{S}_n|} I_{\mathcal{S}_n \neq \emptyset}] \to 0$ implies that $\mathrm{P}(0 < |\mathcal{S}_n| \le A_n) \to 0$ if $\epsilon_n A_n = O(1)$.

In particular, $\text{SENS}(\{f_n\}) \ge \alpha$ if and only if $\mathrm{P}(0 < |\mathcal{S}_n| \le n^{\alpha-\delta}) \to 0$ for all $\delta > 0$.

Exercise 2. Show that $\text{STAB}(\{f_n\}) \le \alpha$ if and only if $\mathrm{P}(|\mathcal{S}_n| \ge n^{\alpha+\delta}) \to 0$ for all $\delta > 0$.

We end this subsection with a brief discussion of general upper bounds on our noise sensitivity and noise stability exponents. We stick for simplicity and without loss of generality to $m_n = n$. We have seen then that 1 is an upper bound for these exponents. On the other hand, the parity function, Example 2, is easily seen to have both these exponents being 1. This question turns out to be much more interesting if we restrict to the important subclass of **monotone** Boolean functions.

Theorem 4.9. *Assume* $f_n : \{0,1\}^n \to \{\pm 1\}$ *be monotone. Then* $\text{STAB}(\{f_n\}) \le 1/2$.

Outline of proof. It is an exercise to check, crucially using the monotonicity, that $I_i(f_n) = |\hat{f_n}(\{i\})|$. It follows that $\sum_i I_i(f_n)^2 \le 1$. The Cauchy-Schwartz inequality now yields that $\sum_i I_i(f_n) \le \sqrt{n}$. The result now follows from Theorem 4.6. □

Remark. The above theorem can be proved (and has been) without the use of Fourier analysis and is known in various contexts.

Answering a question in [4], it was shown in [42] that the above result is optimal by giving a sequence $\{f_n\}$ of monotone functions with $\text{STAB}(\{f_n\}) = 1/2$. By tweaking these examples, one can also obtain a sequence with $\text{SENS}(\{f_n\}) = 1/2$.

5. Critical exponents for percolation

The exact values for critical exponents for percolation on the hexagonal lattice are a crucial ingredient in the study of exceptional times for dynamical percolation and noise sensitivity for crossing probabilities. We therefore briefly describe these.

Let A_R^k be the event (for ordinary percolation) that there are k disjoint monochromatic paths from within distance (say) $2k$ of the origin all of which reach distance R from the origin and such that they are not all of the same color. This is referred to as the k-arm event. Let $A_R^{k,H}$ be the analogous event but where we restrict to the upper half-plane and where the restriction "not all of the same color" may be dropped. This is referred to as the half-plane k-arm event. All these events decay as powers of R and the exact power is called the corresponding critical exponent.

Theorem 5.1. *For the hexagonal lattice, we have*

(i) [38] $\mathrm{P}(A_R^1) = R^{-5/48+o(1)}$

(ii) [53] *For* $k \geq 2$, $\mathrm{P}(A_R^k) = R^{-(k^2-1)/12+o(1)}$

(iii) [53] *For* $k \geq 1$, $\mathrm{P}(A_R^{k,H}) = R^{-k(k+1)/6+o(1)}$

(iv) [53] $\theta(p) = (p - 1/2)^{5/36+o(1)}$.

It was shown by Kesten [31] that (iv) follows from (i) and the case $k = 4$ in (ii). See [43] for a detailed derivation of this.

For dynamical percolation on the hexagonal lattice, A_R^1, A_R^2 and $A_R^{1,H}$ and their exact critical exponents were relevant in the work in [51] while A_R^1, A_R^4, $A_R^{1,H}$ and even some "corner-plane" events and their exact critical exponents were relevant in the work in [14].

We finally mention that the above proofs rely on the conformal invariance of percolation on the hexagonal lattice proved by Smirnov (see [52]) and the convergence of the discrete interface in critical percolation to SLE_6 (see [9] and [52]). SLE originally stood for stochastic Löwner evolution when it was introduced by Schramm in [50] and is presently called the Schramm-Löwner evolution. It has one parameter, usually called κ, and therefore written SLE_κ. As one varies κ, these yield random curves which describe many 2-dimensional critical systems.

6. Exceptional times and positivity of the noise sensitivity exponent for the hexagonal lattice

Considering the question of whether there are exceptional times for dynamical percolation in $\mathbb{Z}^2$, while this was not accomplished in [51], it was proved in this paper that exceptional times do exist for the hexagonal lattice. What allowed the proof to go through for the hexagonal lattice is that various exact critical exponents from critical percolation have been established for the hexagonal lattice. These same critical exponents are expected to hold for $\mathbb{Z}^2$ but have not at this point been established.

Theorem 6.1. [51] *For critical dynamical percolation on the hexagonal lattice, there exist exceptional times of percolation and the Hausdorff dimension of the set of such times is in* $[1/6, 31/36]$.

As far as the noise sensitivity exponent for left to right crossing of an $n \times n$ square, the following was shown.

Theorem 6.2. [51] *Consider the sequence* $\{f_n\}$ *of indicator functions for a left to right crossing of an* $n \times n$ *square in the hexagonal lattice. Then* $\mathrm{SENS}(\{f_n\}) \geq 1/8$. *For the square lattice,* $\mathrm{SENS}(\{f_n\}) > 0$.

This was the first result where one obtained a positive lower bound on $\mathrm{SENS}(\{f_n\})$ for crossing probabilities. One of the key steps in [51] is the following result which gives conditions under which one can obtain some bounds on the Fourier coefficients. We will not give any indication of its proof here. We hope that this result will be applicable in other contexts in order to bound the "level-k" Fourier coefficients.

Theorem 6.3. [51] *Let* $f : \{0,1\}^n \to \mathbb{R}$. *Let* A *be a randomized algorithm determining the value of* f. *This means that* A *examines the input bits of* f *one by one, where the choice of the next bit examined may depend on the values of the bits examined so far and on some additional exterior randomness. The algorithm* A *may of course stop once it knows the output. Let* $J \subseteq \{1, 2, \ldots, n\}$ *be the* (*random*) *set of bits examined by the algorithm.* (*Crucially,* J *need not be all the bits since based on the bits examined at some point, the output might at that point be determined.*) *Set*

$$\delta_A := \sup\{\mathrm{P}(i \in J) : i \in \{1, 2, \ldots, n\}\}.$$

Then, for every $k = 1, 2, \ldots$, *the Fourier coefficients of* f *satisfy*

$$\sum_{|S|=k} \hat{f}(S)^2 \leq \delta_A \, k \, \|f\|^2,$$

where $\|f\|$ *denotes the* L^2 *norm of* f.

We first give an outline of the proof of Theorem 6.2.

Outline of proof. Step 1: Find a randomized algorithm to detect whether there is a left to right crossing of whites such that a fixed hexagon is looked at with probability at most $(1/n)^{1/4+o(1)}$. If we place white hexagons on the right and top sides of the box and black hexagons on the bottom and left sides of the box, we can start at the bottom right and follow the path which always keeps white on the right side and blacks on the left. This is called an **interface**. This interface will end up either hitting the left side before the top, which implies there is a left to right white crossing or it will end up either hitting the top side before the left, which implies there is no left to right white crossing. If we "follow" this interface, revealing hexagon colors as we need to know how the interface develops, this will yield a randomized algorithm which will determine if there is a crossing. In addition, hexagons near the center will be looked at with probability at most $O(1)(1/n)^{1/4+o(1)}$ since to be looked at, one must see the "2-arm" event emanating from that hexagon and one can apply Theorem 5.1(ii). This does not work however for hexagons near the boundary. To get an algorithm which looks at *every* hexagon with the above probability, one does some random modification of the above where one runs two interfaces from a random point on the right side. The argument then requires using the 1-arm half-plane exponent as well.

Step 2: Step 1 and Theorem 6.3 gives us a bound on the sum of the "level-k" Fourier coefficients. We plug that into (4.3) and compute. The $\mathbb{Z}^2$ case is similar but there we do not have the explicit critical exponents at our disposal. □

Remarks. We might hope that one can bring the value 1/8 up to the suggested value of 3/4 by finding better algorithms to which we can apply Theorem 6.3. However, this is not possible. As mentioned in [51], a general inequality of O'Donnell and Servedio or Theorem 6.3 applied in the case $k = 1$ allows us to conclude that any algorithm will have a δ of at least $(1/n)^{1/2+o(1)}$. The existence of such an algorithm would (as above) bring the value of 1/8 up to 1/4 and hence the best this method could yield is a noise sensitivity exponent of 1/4 (unless one improves Theorem 6.3 itself). It is worth pointing out here that an algorithm which is conjecturally better than the one given in the proof of Theorem 6.2 is the one where the hexagon chosen at a given time is the one with the largest influence at the time. This is related to playing random-turn hex; see [48].

We now give an outline of the proof of Theorem 6.1.

Outline of proof. We first explain the existence of exceptional times. We let $X_R := \int_0^1 1_{V_{t,R}}\, dt$ where $V_{t,R}$ is the event that at time t there is an open path from the origin to distance R away. The key step is to show that $\mathrm{E}[X_R^2] \leq O(1)\mathrm{E}[X_R]^2$ (i.e., we use the second moment method) since from here the result is standard. Note the first moment is just $\mathrm{P}(A_R^1)$ from Section 5. The key to getting a good bound on the second moment is to get a good bound on $\mathrm{P}(V_{t,R} \cap V_{0,R})$. Using independence, we see this is at most $\mathrm{P}(V_{0,r})\mathrm{P}(V_{t,r,R} \cap V_{0,r,R})$ where $V_{t,r,R}$ is the event that at time t there is an open path from distance r away to distance R away. We note that looking at our process at times 0 and t is exactly looking at a configuration and its noisy version, which we studied earlier, with ϵ being $1-e^{-t}$. For the second factor, we use (4.3), construct a randomized algorithm for this event with a good δ (which is somewhat harder than in Theorem 6.2) and apply Theorem 6.3 to bound the relevant Fourier coefficients. The rest is algebra.

For the Hausdorff dimension, the lower bound is obtained using the calculation in the first paragraph together with Frostman's Theorem. The upper bound is easier; we use the method of proof of Theorem 2.5 together with Theorem 5.1(iv). □

In [51], other types of events are also looked at such as k-arm events in wedges and cones and upper and lower bounds on the Hausdorf dimension of sets of exceptional times are obtained. These upper and lower bounds are however never matching.

7. The exact Hausdorff dimension of exceptional times and the exact noise sensitivity exponent for the hexagonal lattice

In [14], two of the main results (among many others) were computing the exact noise sensitivity exponent for the crossing of an $n \times n$ box in the hexagonal lattice and the exact Hausdorff dimension of the set of exceptional times for dynamical percolation on the hexagonal lattice. The latter number is the upper bound obtained in Theorem 6.1.

Theorem 7.1. [14] *For critical dynamical percolation on the hexagonal lattice, the Hausdorff dimension of the set of exceptional times is* $31/36$. *In addition, for critical dynamical percolation on the square lattice, there exist exceptional times of percolation.*

Theorem 7.2. [14] *Consider the sequence* $\{f_n\}$ *of indicator functions for a left to right crossing of an* $n \times n$ *square in the hexagonal lattice. Then*

$$\mathrm{STAB}(\{f_n\}) = \mathrm{SENS}(\{f_n\}) = 3/4.$$

(While we will not spell these out in detail, for the square lattice, analogous results are obtained which relate $\mathrm{STAB}(\{f_n\})$ and $\mathrm{SENS}(\{f_n\})$ with the expected number of pivotal edges in large boxes.)

The value $31/36$ was the conjectured value in [51] and the suggested value of $3/4$ was explained earlier. The improvements here over Theorems 6.1 and 6.2 are due almost exclusively to providing much sharper results concerning the Fourier spectrum, both for crossing probabilities and an annulus type event which is the relevant event for studying exceptional times. These arguments do not for example use Theorem 6.3 or any variant of this. This analysis is very long and intricate and so I will only say a few words about it and even for that I will stick to Theorem 7.2.

(*Very vague*) *Outline of proof.* The much easier direction is to show that $\mathrm{STAB}(\{f_n\}) \leq 3/4$. For a hexagon to be pivotal, there have to be, starting from that hexagon, "open paths" to the left and right sides of the box and "closed paths" to the top and bottom sides of the box. It is known that this has the same exponent as the 4-arm event. Looking at points far from the boundary and using Theorem 5.1(ii), we obtain that $I(f_n) \geq n^{3/4+o(1)}$. In [14], it is shown that the boundary contributions can be controlled so that we indeed have $I(f_n) = n^{3/4+o(1)}$. Now Theorem 4.6 finishes the argument.

The proof that $\mathrm{SENS}(\{f_n\}) \geq 3/4$ is significantly more difficult. By Theorem 4.8, we need to prove (4.6) and so we need to obtain upper bounds on the lower tail of the distribution of $|\mathcal{S}_n|$. Of course, we only care about the distribution of $|\mathcal{S}_n|$ rather than the distribution of $\mathcal{S}_n$. However, it turns out that in order to study and analyze the former, it is crucial to study the latter, which has much more structure and therefore more amenable to analysis. A first key step is for general Boolean functions and gives upper bounds on

$$\mathrm{P}(\mathcal{S} \cap B \neq \emptyset = \mathcal{S} \cap W), \tag{7.1}$$

where B and W are disjoint subsets of the domain variables $1, \ldots, n_m$, in terms of a more general notion of pivotality. While one needs such a result for all W, looking at the two special cases where W is $\emptyset$ or B^c illustrates well this relationship. This general result gives in the context of percolation that if B is a connected set of hexagons, $\mathrm{P}(\mathcal{S} \cap B \neq \emptyset)$ is at most 4 times the probability of having 4 alternating arms from the set B out to the boundary of our $n \times n$ box and $\mathrm{P}(\emptyset \neq \mathcal{S} \subseteq B)$ is at most 4 times the previous probability squared. This starts to get the ball rolling as it relates the difficult spectral measure to things that are a little bit more concrete.

Now let $g_r := r^2 \alpha_4(r)$ which is close to the expected number of pivotals for a left to right crossing in an $r \times r$ box or equivalently the expected size of the spectral sample for this event. This grows to ∞ with r. One shows that

$$\mathrm{P}(|\mathcal{S}_n| < g_r, \mathcal{S}_n \neq \emptyset) \asymp (n/r)^2 (\alpha_4(n)/\alpha_4(r))^2. \tag{7.2}$$

It is not hard to show, using the fact that the 4 arm exponent is $5/4$, that (4.6) holds and that one can take δ to be $\gamma^{3/2+\epsilon}$ for any fixed $\epsilon > 0$. While we want the upper bound, it is instructive to see how the lower bound is obtained which is as follows. We break the $n \times n$ square into about $(n/r)^2$ $r \times r$ squares. It turns out that in the percolation context and with B being an $r \times r$ square and $W = B^c$, the upper bound on (7.1) is shown to also be a lower bound (up to constants) and so for each $r \times r$ square B, the probability that the spectrum is nonempty and sits inside B can been shown to be at least $\Omega(1)(\alpha_4(n)/\alpha_4(r))^2$. Next it is shown that, conditioned on the spectrum intersecting such a box and no others, there is a uniform lower bound on the probability that the spectral size within that box is at most $O(1)g_r$. Since, as we vary the $r \times r$ square, we obtain $(n/r)^2$ disjoint events, we obtain the (much easier) lower bound of (7.2).

For the upper bound (which is much harder), we again break the $n \times n$ square as above and look at the number of $r \times r$ squares which intersect the spectral sample. Call this number $X_{n,r}$. Using a very difficult geometric induction argument, it is shown that

$$\mathrm{P}(X_{n,r} = k) \leq g(k)(n/r)^2 (\alpha_4(n)/\alpha_4(r))^2$$

where $g(k)$ grows slower than exponentially (but faster than any polynomial). It turns out that there is a "converse" of what we wrote in the previous paragraph which is that conditioned on the spectrum touching a box, there is uniform lower bound on the probability that the size of the spectrum is at least $\Omega(1)g_r$. This involves quite difficult percolation arguments. Given $X_{n,r} = k$, if it were the case that the sizes of the spectrum in the k different $r \times r$ boxes which the spectrum hits were independent, then the probability that all of them have size less than $\Omega(1)g_r$ would be at most c^k for some $c < 1$. Since $\sum_k g(k)c^k < \infty$, we would be done. The problem is, under this conditioning, the spectrum in the different boxes are not independent and have a very complicated dependency structure. It is however shown that, while it is difficult to deal with the spectrum in one box conditioned on its behavior elsewhere, it is possible to deal with the spectrum in one box conditioned on it hitting that box and not intersecting some other fixed set. This together with a novel type of large deviations argument allows us to carry out the upper bound in (7.2). □

In proving the above, other interesting results are obtained. For example, it is shown for percolation on $\mathbb{Z}^2$ that rerandomizing a small portion of only the vertical edges is sufficient for the correlation to go to 0. This result suggests new interesting directions concerning noise sensitivity for Boolean functions when only some of the bits are rerandomized.

8. Sensitivity of the infinite cluster in critical percolation

Except in the second part of Theorem 2.4, all results in this paper so far dealt with graphs which do not percolate at criticality. It turns out that if we deal with graphs which do percolate at criticality and ask if there are exceptional times of *nonpercolation*, the structure of the problem is quite different. In addition, it seems to me that the set of exceptional times in this case might have similar structure to the set of "fast points" for Brownian motion; see the discussion at the end of Section 3.

In [47], among other things, a fairly detailed study of this question was made for spherically symmetric trees. Special cases of the two main results of that paper are the following. In this section (only), we are dropping the assumption of homogeneous edge probabilities but will assume all edge probabilities are bounded away from 0 and 1. The definition of w_n from Section 3 should be modified in the obvious way.

Theorem 8.1. *Consider a spherically symmetric tree with spherically symmetric edge probabilities (meaning all edges at a given level have the same retention probability). Let w_n be as in Theorem* 3.1.

(i) *If*

$$\lim_n \frac{w_n}{n(\log n)^\alpha} = \infty$$

for some $\alpha > 2$, then there are no exceptional times of nonpercolation.

(ii) *If*

$$w_n \asymp n(\log n)^\alpha$$

for some $1 < \alpha \leq 2$, then there are exceptional times of nonpercolation.

Note that in both of these regimes, Theorem 3.1 *tells us that there is percolation at a fixed time.*

Remarks.

(1) To see a concrete example, if we have a tree with $|T_n| \asymp 2^n n(\log n)^\alpha$ and $p = 1/2$ for all edges, then if $\alpha > 2$, we are in case (i) while if $\alpha \leq 2$, we are in case (ii). (Note Lyons' theorem tells us that $p_c = 1/2$ in these cases.)

(2) The theorem implies that if $w_n \asymp n^\alpha$ with $\alpha > 1$, then there are no exceptional times of nonpercolation, while note that if $w_n \asymp n$, then Theorem 3.1

implies that there is no percolation at a fixed time. Hence, if we only look at the case where $w_n \asymp n^\alpha$ for some $\alpha \geq 1$, we do not see the dichotomy (within the regime where we percolate at criticality) that we are after but rather we see the much more abrupt transition from not percolating at a fixed time to percolating at all times. Rather, Theorem 8.1 tells us that we need to look at a "finer logarithmic scale" to see this "phase transition" of changing from percolating at a fixed time but having exceptional times (of nonpercolation) to percolating at all times.

Interestingly, it turns out that even within the regime where there are no exceptional times of nonpercolation, there are still two very distinct dynamical behaviors of the process, yielding another phase transition.

Theorem 8.2. *Consider a spherically symmetric tree T of bounded degree and let w_n be as in the previous result.*

(i) *When $\sum_{n=1}^{\infty} n\, w_n^{-1} < \infty$, a.s. the set of times $t \in [0,1]$ at which the root percolates has finitely many connected components.* (*This holds for example if $w_n \asymp n^\theta$ with $\theta > 2$ as well as for supercritical percolation on a homogeneous tree.*)

(ii) *If $w_n \asymp n^\theta$, where $1 < \theta < 2$, then with positive probability the set of times $t \in [0,1]$ at which the root percolates has infinitely many connected components. The same occurs if $w_n \asymp n(\log n)^\alpha$ for any $\alpha > 1$.*

Remarks.

(1) If $w_n \asymp n^2$, we do not know the answer. A first moment calculation suggests that there should be infinitely many connected components with positive probability but the needed inequality for a successful second moment argument fails.

(2) It is easy to show that for any graph, if there are exceptional times of nonpercolation, then the set of times $t \in [0,1]$ at which a fixed vertex percolates is totally disconnected and hence has infinitely many connected components with positive probability.

(3) We will not indicate here any proofs but we will mention one word about Theorem 8.2 since the critical exponent of 2 there derives from a difference in the ordinary model in these two regimes. Namely, the expected number of pivotal edges for the event that the root percolates is infinite for $\theta \leq 2$ but finite for $\theta > 2$.

9. Dynamical percolation and the incipient infinite cluster

9.1. The incipient infinite cluster

We know that on $\mathbb{Z}^2$ and on the hexagonal lattice, there is no infinite cluster at p_c. Nonetheless, physicists and others have tried to talk about the "infinite cluster on $\mathbb{Z}^2$ containing the origin at criticality". This was made sense of by Kesten in

[29] where the following result was obtained. Λ_n is the box of size n centered at the origin.

Theorem 9.1. [29] *The limiting measures*

$$\lim_{p\downarrow 1/2} \mathrm{P}_p(\cdot \mid 0 \leftrightarrow \infty) \quad \textit{and} \quad \lim_{n\to\infty} \mathrm{P}_{1/2}(\cdot \mid 0 \leftrightarrow \partial\Lambda_n)$$

both exist and are equal.

This limiting measure is referred to as the **incipient infinite cluster**. Properties of this were also obtained and furthermore, in [30], Kesten showed that random walk on the incipient infinite cluster is subdiffusive.

9.2. The incipient infinite cluster and dynamical percolation

It was asked quite early on whether the configuration for dynamical percolation at a (properly chosen) exceptional time at which the origin percolates (assuming there are such exceptional times) should have the same distribution as the incipient infinite cluster of the previous subsection. See the discussion concerning this question in [17]. This question was answered in a very satisfactory manner by Hammond, Pete and Schramm in [20], a paper which is in the process of being written up. We sketch here a part of what was done in this paper.

The first key step in being able to "find" the incipient infinite cluster inside of dynamical percolation is to be able to define a *local time* for when the origin is percolating. There are two different approaches used to define a local time.

The first approach is as follows. Let $A_{R,t}$ be the event that at time t there is an open path from the origin to distance R away and let $\mathcal{T}_R$ be the random set of times at which $A_{R,t}$ occurs. Define a random measure μ_R on $\mathbb{R}$ by

$$\mu_R := \frac{1}{\alpha_1(R)} \mathcal{L}_{\mathcal{T}_R}$$

where $\mathcal{L}_F$ refers to Lebesgue measure restricted to the set F and $\alpha_1(R) = \mathrm{P}(A_{R,t})$. Clearly, $\mathrm{E}[\mu_R([a,b])] = b - a$. We then let R tend to infinity.

The second approach (which turns out to be equivalent) is as follows and is closer in spirit to Kesten's original definition of the incipient infinite cluster. Consider ordinary percolation and let S be a collection of hexagons and let ω^S be the percolation realization restricted to S. We want to measure in some sense how much ω^S "helps percolation to infinity". This is made precise by the limit

$$\lim_{R\to\infty} \frac{\mathrm{P}(A_R \mid \omega^S)}{\mathrm{P}(A_R)}$$

which is easily shown to exist using Theorem 9.1.

Calling this limit $f(\omega^S)$, let $M_r(\omega) := f(\omega^{B_r})$ where B_r is the ball of radius r around the origin. Finally let

$$\nu_r([a,b]) := \int_a^b M_r(\omega_s)ds.$$

Theorem 9.2. [20]

(i) *For all $a < b$, $\mu_R([a,b])$ converges a.s. and in L^2 to a limit as R goes to ∞, which we call $\mu_\infty([a,b])$.*

(ii) *For all $a < b$, $\nu_r([a,b])$ converges a.s. and in L^2 to a limit as r goes to ∞, which we call $\nu_\infty([a,b])$.*

(iii) *The two above limits are a.s. the same.*

Clearly the limiting measure μ_∞ is supported on the set of exceptional times at which the origin percolates (the latter known to be a nonempty closed set). It is not known if the support of the measure is exactly the set of exceptional times. It is explained that $\mu_R([a,b])$ is a martingale which implies its a.s. convergence. Using estimates from [51], one can see it is also L^2 bounded which gives the L^2 convergence. The convergence in L^2 guarantees that the limiting measure is nonzero. It is also shown that $\nu_r([a,b])$ is a martingale.

The final result links up the incipient infinite cluster with dynamical percolation.

Theorem 9.3. [20] *Consider the random measure μ_∞ on $\mathbb{R}$ above and let X be a Poisson process on $\mathbb{R}$ with "intensity measure" μ_∞. Then the distribution of ω_0 given $0 \in X$ has the same distribution as the incipient infinite cluster.*

It is not so hard to make sense of the conditioning $0 \in X$ even if this event has probability 0; see Chapter 11 of [27]. There are a number of other results in this paper which we do not detail here.

10. The scaling limit of planar dynamical percolation

Before discussing the scaling limit of dynamical percolation, it is necessary to first discuss the scaling limit of ordinary percolation. There is a lot to be said here and this section will be somewhat less precise than the earlier sections. Since even the formulations can be quite technical, I will be, unlike in the previous sections, "cheating" in various places.

Even before we state the scaling limit of percolation, we need to first briefly explain the concept of conformal invariance and Cardy's formula. Let Ω be a simply connected open domain in the plane and let A, B, C and D be 4 points on the boundary of Ω in clockwise order. Scale a 2-dimensional lattice, such as $\mathbb{Z}^2$ or the hexagonal lattice, by $1/n$ and perform critical percolation on this scaled lattice. Let $\mathrm{P}(\Omega, A, B, C, D, n)$ denote the probability that, in the $1/n$ scaled hexagonal lattice, there is a white path of hexagons inside Ω going from the boundary of Ω between A and B to the boundary of Ω between C and D. The first half of the following conjecture was stated in [36] and attributed to Michael Aizenman. The second half of the conjecture is due to Cardy [11].

Conjecture 10.1.

(i) *For all Ω and A, B, C and D as above,*

$$\mathrm{P}(\Omega, A, B, C, D, \infty) := \lim_{n\to\infty} \mathrm{P}(\Omega, A, B, C, D, n)$$

exists and is conformally invariant in the sense that if f is a conformal mapping, then $\mathrm{P}(\Omega, A, B, C, D, \infty) = \mathrm{P}(f(\Omega), f(A), f(B), f(C), f(D), \infty)$.

(ii) *There is an explicit formula* (*not stated here*) *for* $\mathrm{P}(\Omega, A, B, C, D, \infty)$*, called Cardy's formula, when Ω is a rectangle and A, B, C and D are the* 4 *corner points.* (*Since every Ω, A, B, C and D can be mapped to a unique such rectangle* (*with A, B, C and D going to the* 4 *corner points*)*, this would specify the above limit in general assuming conformal invariance.*)

Cardy's formula is quite complicated involving hypergeometric functions but Lennart Carleson realized that assuming conformal invariance, there is a nicer set of "representing" domains with four specified points for which the limit has a much simpler form. Namely, if Ω is an equilateral triangle (with side lengths 1), A, B and C the three corner points and D (on the line between C and A) having distance x from C, then the above probability would just be x. Using Carleson's reformulation of Cardy's formula, Smirnov proved the above conjecture for the hexagonal lattice.

Theorem 10.2. [52] *For the hexagonal lattice, both* (i) *and* (ii) *of the above conjecture are true.*

This conjecture is also believed to hold on $\mathbb{Z}^2$ but is not (yet) proved in that case. In [50], Schramm described what the interfaces between whites and blacks should be as the lattice spacing goes to 0, assuming conformal invariance. In the appropriate formulation, it should be an SLE_6 curve. Smirnov [52] proved this convergence for one interface and Camia and Newman [9] proved a "full scaling limit", which is a description of the behavior of all the interfaces together. The critical exponents described in Section 5 are proved by exploiting the SLE description of the interfaces. All of the above is described well in [54].

It turns out, in order to obtain a scaling limit of dynamical percolation, a different description, due to Schramm and Smirnov, of the scaling limit of ordinary percolation is preferable. A quad Q is a subset of the plane homeomorphic to a disk together with its boundary partitioned into 4 continuous pieces. A configuration for the scaling limit is described by a family of 0-1 random variables X_Q indexed by the quads Q where $X_Q = 1$ means there is a crossing from the first to the third boundary piece of Q using white hexagons. Equivalently, an element of the state space is a collection of quads (satisfying a certain necessary monotonicity condition and suitably measurable) where the collection of quads represents the quads which are "crossed". An important advantage of this state space, which we denote by $\mathcal{S}$, is that it is a compact metrizable space. This avoids the need for tightness arguments. If we perform percolation on the $1/n$ scaled hexagonal lattice, the X_Q's are well defined random variables and so we obtain a probability measure μ_n on $\mathcal{S}$. It turns out that the sequence of probability measures μ_n have a limit μ_∞, which we call the scaling limit. Note that for each n, there is essentially a 1-1 correspondence between percolation realizations on the $1/n$ scaled lattice and elements in $\mathcal{S}$. In the limit however, we can not talk anymore about percolation configurations but rather the only information left is which quads are "crossed". (The latter is sort of the "macroscopic" information.)

We now move to a scaling limit for dynamical percolation. Consider dynamical percolation on the $1/n$ scaled hexagonal lattice. We can think of our dynamical percolation on this scaled lattice as a process $\{\eta^n(t)\}_{t\in\mathbb{R}}$ taking values in $\mathcal{S}$. We can try to let n go to infinity, in which case, the marginal for each t will certainly go to μ_∞. However due to the noise sensitivity of percolation, for any fixed $s < t$ and any quad Q, the crossing of Q at time s and at time t will become asymptotically independent as $n \to \infty$ which implies that the processes $\{\eta^n(t)\}_{t\in\mathbb{R}}$ converge to something which is "independent copies of μ_∞ for all different times". This is clearly not what we want. In order to obtain a nontrivial limit, we should slow down time by a factor of $1/(n^2\alpha_4(n))$ where $\alpha_4(n)$ is explained in Section 7.

Theorem 10.3. [15] *Let* $\sigma_t^n := \eta^n_{tn^{-2}\alpha_4^{-1}(n)}$. *Then* $\{\sigma^n(t)\}_{t\in\mathbb{R}}$ *converges in law to a process* $\{\sigma^\infty(t)\}_{t\in\mathbb{R}}$ *under the topology of local uniform convergence. Moreover, the limiting process is Markovian* (*which is not a priori obvious at all*).

We explain now why the time scaling is as above. Consider the quad corresponding to the unit square together with its four sides. The expected number of pivotals for a left to right crossing of this quad on the $1/n$ scaled lattice is $\asymp n^2\alpha_4(n)$. Next, the above time scaling updates each hexagon in unit time with probability about $1/n^2\alpha_4(n)$ and therefore, by the above, updates on average order 1 pivotals. It follows from the main result in [14] that if we scale "faster" than this, you get a limit which is independent at different times and if we scale "slower" than this, you get a limit which is constant in time. This scaling is such that the correlation between this event occurring at time 0 and occurring at time 1 stays strictly between 0 and 1, which is something we of course want.

In the above paper, there is another model which is studied called near-critical percolation for which a scaling limit is proved in a similar fashion. Near-critical percolation can be thought of as a version of dynamical percolation, where sites are flipping only in one direction and hence after time t, the density of open sites is about $1/2 + t/(n^2\alpha_4(n))$.

To prove Theorem 10.3, there are two key steps. The first key step is a stability type result. In very vague terms, it says that if we look at the configuration at time 0, ignore the evolution of the hexagons which are initially pivotal only for crossings which occur at macroscopic scale at most ρ (which should be thought of as much larger than the lattice spacing) but observe the evolution of the hexagons which are initially pivotal for crossings which occur at macroscopic scale larger than ρ, then we can still predict quite well (i.e., with high probability if ρ is small) how the crossings evolve on macroscopic scale 1.

Since we cannot see the individual hexagons which are initially pivotal for crossings at scale at least ρ in the scaling limit, in order for this stability result to be useful, we need at least that the number of such hexagons in a given region can be read off from the scaling limit, so that we know the rate with which macroscopic crossings are changing. This is obtained by the second key step. For every n, look at ordinary percolation on the $1/n$ scaled lattice and consider the set of hexagons from which emanate 4 paths of alternating colors to macroscopic-distance ρ away. (These are the hexagons which, if flipped, change the state of a crossing at scale ρ.) Let ν_n^ρ be counting measure on such points divided by $n^2\alpha_4(n)$. This scaling is such that the expected measure of the unit square is of order 1 (as $n \to \infty$ if ρ is fixed).

Theorem 10.4. [15] *For all ρ, (ω_n, ν_n^ρ) converges, as $n \to \infty$, to a limit $(\omega_\infty, \nu_\infty^\rho)$ and ν_∞^ρ is a measurable function of ω_∞.*

The argument of this result is very difficult. I say however a word on how this gives a nice construction of the scaling limit of dynamical percolation. The idea is that one constructs a dynamical percolation realization by first taking a realization from the $t = 0$ scaling limit (or equivalently ω_∞ above), looks at ν_∞^ρ (which requires no further randomness because ν_∞^ρ has been proved to be a function of ω_∞) and then builds an appropriate Poisson point process over the random measure $\nu_\infty^\rho \times dt$ which will be used to dictate when the different crossings of quads of scale at least ρ will change their state. This is done for all ρ and the stability result is needed to ensure that the process described in this way is in fact the scaling limit of dynamical percolation. The idea that dynamical percolation (and near-critical percolation) could possibly be built up in this manner from a Poisson process over a random measure was suggested by Camia, Fontes and Newman in the paper [10].

It is also shown in [15] that a type of "conformal covariance" for the above measures ν_∞^ρ holds but this will not be detailed here. The argument for Theorem 10.4 also proves the existence of natural time-parametrizations for the SLE$_6$ and SLE$_{8/3}$ curves, a question studied in [37] for general SLE$_\kappa$.

11. Dynamical percolation for interacting particle systems

In [7], certain interacting particle systems were studied and the question of whether there are exceptional times at which the percolation structure looks different from that at a fixed time was investigated. The two systems studied were the contact process and the Glauber dynamics for the Ising model. Most of the results dealt with noncritical cases but since the dynamics are not independent, a good deal more work was needed compared to the easy proof of Proposition 2.1.

However, there was one very interesting case left open in this work which was the following. Consider the Glauber dynamics for the critical 2-dimensional Ising model. First note that 2 dimensions is special for the Ising model compared to higher dimensions since it is known that for $\mathbb{Z}^2$ the critical value for phase transition is the same as the critical value for when percolation occurs. The Ising model does not percolate in 2 dimensions at the critical value and in view of Theorem 6.1, it is natural to ask if there are exceptional times of percolation for the Glauber dynamics. During a visit to Microsoft, Gábor Pete suggested to me that this might not be the case since the scaling limit of the Ising model was conjectured to be SLE_3, which unlike SLE_6 (the scaling limit of percolation), does not hit itself. This yields that there should not be so many pivotal sites which suggests no exceptional times. A few months later in Park City, Christophe Garban sketched out a “1st moment argument” for me which would give no exceptional times under certain assumptions.

In any case, whether or not one can actually prove this, it now seems clear that there are no exceptional times for the Glauber dynamics. This seems to be a beautiful example of how the qualitative difference in the behavior of the two respective SLE scaling limits (self-intersections versus not) yields a fundamental difference between the natural dynamics on ordinary percolation and the natural dynamics on the Ising model concerning whether there are exceptional times of percolation or not.

Acknowledgments

I would like to thank Erik Broman, Christophe Garban, Olle Häggström, Alan Hammond, Russell Lyons and Gábor Pete for very many helpful comments on earlier versions of the manuscript. I appreciate in particular Christophe Garban, Alan Hammond and Gábor Pete explaining to me their unpublished work. A part of this work was done while visiting the Mittag-Leffler Institute in Djursholm, Sweden, the same institute where dynamical percolation was started.

References

[1] Adelman, O., Burdzy, K. and Pemantle, R. Sets avoided by Brownian motion. *Ann. Probab.* **26**, (1998), 429–464.

[2] Aizenman, M., Kesten, H. and Newman, C.M. Uniqueness of the infinite cluster and continuity of connectivity functions for short- and long-range percolation. *Comm. Math. Phys.* **111**, (1987), 505–532.

[3] Benjamini, I., Häggström, O., Peres, Y. and Steif, J. Which properties of a random sequence are dynamically sensitive? *Ann. Probab.* **31**, (2003), 1–34.

[4] Benjamini, I., Kalai, G. and Schramm, O. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.* **90**, (1999), 5–43.

[5] Benjamini, I. and Schramm, O. Exceptional planes of percolation. *Probab. Theory Related Fields* **111**, (1998), 551–564.

[6] Berg, J. van den, Meester, R. and White, D.G. Dynamic Boolean models. *Stochastic Process. Appl.* **69**, (1997), 247–257.

[7] Broman, E.I. and Steif, J.E. Dynamical Stability of Percolation for Some Interacting Particle Systems and ϵ-Movability. *Ann. Probab.* **34**, (2006), 539–576.

[8] Burton, R. and Keane, M. Density and uniqueness in percolation. *Comm. Math. Phys.* **121**, (1989), 501–505.

[9] Camia, F. and Newman, C.M. Two-dimensional critical percolation: the full scaling limit, *Comm. Math. Phys.* **268**, (2006), 1–38.

[10] Camia, F., Fontes, L.R.G. and Newman, C.M. Two-dimensional scaling limits via marked nonsimple loops. *Bull. Braz. Math. Soc.* (*N.S.*) **37**, (2006), 537–559.

[11] Cardy, J.L. Critical percolation in finite geometries. *J. Phys. A* **25**, (1992), L201–L206.

[12] Evans, S.N. Local properties of Levy processes on a totally disconnected group. *J. Theoret. Probab.* **2**, (1989), 209–259.

[13] Garban, C. Oded Schramm's contributions to noise sensitivity (preliminary title), in preparation.

[14] Garban, C., Pete, G. and Schramm, O. The Fourier Spectrum of Critical Percolation, preprint, arXiv:0803.3750[math:PR].

[15] Garban, C., Pete, G. and Schramm, O. Scaling limit of near-critical and dynamical percolation, in preparation.

[16] Grimmett, G. *Percolation.* Second edition, Springer-Verlag, (1999), New York.

[17] Häggström, O. Dynamical percolation: early results and open problems. *Microsurveys in discrete probability* (Princeton, NJ, 1997), 59–74, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 41, Amer. Math. Soc., Providence, RI, 1998.

[18] Häggström, O. and Pemantle, R. On near-critical and dynamical percolation in the tree case. Statistical physics methods in discrete probability, combinatorics, and theoretical computer science (Princeton, NJ, 1997). *Random Structures Algorithms* **15**, (1999), 311–318.

[19] Häggström, O., Peres, Y. and Steif, J.E. Dynamical percolation. *Ann. Inst. Henri Poincaré, Probab. et Stat.* **33**, (1997), 497–528.

[20] Hammond, A., Pete, G. and Schramm, O. Local time for exceptional dynamical percolation, and the incipient infinite cluster, in preparation.

[21] Hara, T. and Slade, G. (1994) Mean field behavior and the lace expansion, in *Probability Theory and Phase Transitions*, (ed. G. Grimmett), Proceedings of the NATO ASI meeting in Cambridge 1993, Kluwer.

[22] Harris, T.E. A lower bound on the critical probability in a certain percolation process. *Proc. Cambridge Phil. Soc.* **56**, (1960), 13–20.

[23] Jonasson, J. Dynamical circle covering with homogeneous Poisson updating *Statist. Probab. Lett.*, to appear.

[24] Jonasson, J. and Steif, J.E. Dynamical models for circle covering: Brownian motion and Poisson updating. *Ann. Probab.* **36**, (2008), 739–764.

[25] Kahn, J., Kalai, G. and Linial, N. The influence of variables on boolean functions. 29*th Annual Symposium on Foundations of Computer Science*, (1988), 68–80.

[26] Kalai, G. and Safra, S. Threshold phenomena and influence: perspectives from mathematics, computer science, and economics. Computational complexity and statistical physics, 25–60, St. Fe Inst. Stud. Sci. Complex., Oxford Univ. Press, New York, 2006.

[27] Kallenberg, O. *Foundations of modern probability.* Second edition. Probability and its Applications (New York). Springer-Verlag, New York, 2002.

[28] Kesten, H. The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$. *Comm. Math. Phys.* **74**, (1980), 41–59.

[29] Kesten, H. The incipient infinite cluster in two-dimensional percolation. *Probab. Theory Related Fields* **73**, (1986), 369–394.

[30] Kesten, H. Subdiffusive behavior of random walk on a random cluster. *Ann. Inst. Henri Poincaré, Probab. et Stat.* **22**, (1986), 425–487.

[31] Kesten, H. Scaling relations for 2D-percolation. *Commun. Math. Phys.* **109**, (1987), 109–156.

[32] Kesten, H. and Zhang, Y. Strict inequalities for some critical exponents in 2D-percolation. *J. Statist. Phys.* **46**, (1987), 1031–1055.

[33] Khoshnevisan D. Dynamical percolation on general trees. *Probab. Theory Related Fields* **140**, (2008), 169–193.

[34] Khoshnevisan, D. Multiparameter processes. An introduction to random fields. Springer Monographs in Mathematics. Springer-Verlag, New York, 2002.

[35] Khoshnevisan, D., Peres, Y. and Xiao, Y. Limsup random fractals. *Electron. J. Probab.* **5**, (2000), no. 5, 24 pp. (electronic).

[36] Langlands, R., Pouliot, P. and Saint-Aubin, Y. Conformal invariance in two-dimensional percolation. *Bull. Amer. Math. Soc.* (*N.S.*) **30**, (1994), 1–61.

[37] Lawler, G. Dimension and natural parametrization for SLE curves.

[38] Lawler, G., Schramm, O. and Werner, W. One-arm exponent for critical 2D percolation. *Electron. J. Probab.* **7**, (2002), no. 2, 13 pp. (electronic).

[39] Lyons, R. Random walks and percolation on trees. *Ann. Probab.* **18**, (1990), 931–958.

[40] Lyons, R. Random walks, capacity, and percolation on trees. *Ann. Probab.* **20**, (1992), 2043–2088.

[41] Lyons, R. with Peres, Y. (2008). *Probability on Trees and Networks.* Cambridge University Press. In preparation. Current version available at http://mypage.iu.edu/˜rdlyons/.

[42] Mossel, E. and O'Donnell, R. On the noise sensitivity of monotone functions. *Random Structures Algorithms* **23**, (2003), 333–350.

[43] Nolin, P. Near-critical percolation in two dimensions. *Electron. J. Probab.* **13**, (2008), no. 55, 1562–1623.

[44] O'Donnell, R. Computational applications of noise sensitivity, Ph.D. thesis, MIT (2003). Version available at http://www.cs.cmu.edu/∼odonnell/.

[45] Peres, Y. Intersection-equivalence of Brownian paths and certain branching processes. *Commun. Math. Phys.* **177**, (1996), 417–434.

[46] Peres, Y. Remarks on intersection-equivalence and capacity-equivalence. *Ann. Inst. Henri Poincaré* (*Physique théorique*) **64**, (1996), 339–347.

[47] Peres, Y., Schramm, O. and Steif, J.E. Dynamical sensitivity of the infinite cluster in critical percolation. *Ann. Inst. Henri Poincaré, Probab. et Stat.*, to appear.

[48] Peres, Y., Schramm, O., Sheffield, S. and Wilson, D.B. Random-turn hex and other selection games. *Amer. Math. Monthly* **114**, (2007), 373–387.

[49] Peres, Y. and Steif, J.E. The number of infinite clusters in dynamical percolation. *Probab. Theory Related Fields* **111**, (1998), 141–165.

[50] Schramm, O. Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.* **118**, (2000), 221–288.

[51] Schramm, O. and Steif, J.E. Quantitative noise sensitivity and exceptional times for percolation. *Ann. Math.*, to appear.

[52] Smirnov, S. Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *C.R. Acad. Sci. Paris Sér. I Math.* **333**, (2001), 239–244.

[53] Smirnov, S. and Werner, W. Critical exponents for two-dimensional percolation. *Math. Res. Lett.* **8**, (2001), 729–744.

[54] Werner, W. Lectures on two-dimensional critical percolation. IAS Park City Graduate Summer School, 2007, arXiv:0710.0856[math:PR].

Jeffrey E. Steif
Mathematical Sciences
Chalmers University of Technology

and

Mathematical Sciences
Göteborg University
SE-41296 Gothenburg, Sweden
e-mail: `steif@chalmers.se`

Progress in Probability, Vol. 61, 175–196
© 2009 Birkhäuser Verlag Basel/Switzerland

Measure-valued Processes, Self-similarity and Flickering Random Measures

Jochen Blath

Abstract. In this article, we discuss two of the main prototypes of measure-valued processes, namely the classical Fleming-Viot and Dawson-Watanabe processes, and some of their recent generalizations. In particular, we show how the so-called lookdown construction of Donnelly and Kurtz can be used to reveal interesting structural- and path-properties of the (generalized) processes in the case when the underlying motion and branching mechanisms satisfy certain self-similarity properties. As applications of the method, we first discuss the notion of a 'flickering random measure', and then conclude with remarks about properties of the support of general, and in particular Beta-, Fleming-Viot processes.

Mathematics Subject Classification (2000). Primary: 60G57;
Secondary: 60G17.

Keywords. Generalized Fleming-Viot process, flickering random measures, super-Brownian motion, Dawson-Watanabe process, measure-valued diffusion, coalescent, lookdown construction, wandering random measure, Neveu superprocess.

1. Introduction

Measure-valued processes are stochastic processes that typically arise as limits of empirical measures of interacting stochastic particle systems. The main ingredients of such systems are the underlying motion (resp. mutation) of particles and their reproduction mechanism.

While such models have interesting applications, e.g., in mathematical genetics and population biology, they also reveal a rich mathematical structure. Sophisticated methods have been introduced to construct and explore these processes, e.g., Bertoin and Le Gall's flow of bridges [1], Le Gall's Brownian Snake [25] resp. Le Gall's and Le Jan's infinite variance snake [26], Donnelly and Kurtz' (modified) lookdown construction [12], [13] and, very recently, Kurtz and Rodriguez' Poisson-type construction [24].

In this article, we focus on the interplay and the properties of two classical prototypes of measure-valued processes, namely the Fleming-Viot and the Dawson-Watanabe process, and their (recent) generalizations. We will show how the lookdown-construction of Donnelly and Kurtz can be used to clarify the relationship between both classes of processes and provides insight into some of their path properties.

2. Two classical measure-valued processes as universal limits

2.1. The Dawson-Watanabe process

This process, also known as "super-Brownian motion" was introduced in 1968 by Shinzo Watanabe [43] as a limit of branching Brownian motions.

Indeed, let $N \in \mathbb{N}$ and consider a system of N particles whose initial positions (i.e., at time $t = 0$) are i.i.d. distributed according to some probability measure μ on $\mathbb{R}^d$. We assume that the *underlying motion* is Brownian, i.e., during its lifetime, each particle moves around in $\mathbb{R}^d$ according to a standard Brownian motion. Moreover, we assume the following *reproduction* resp. *branching mechanism*: Each particle has an exponentially distributed lifetime with parameter $\gamma > 0$. When it dies, either two or zero offspring (each with probability 1/2) are left behind at the location where the particle died, which then again evolve independently according to the same mechanism as their parent. Note that our special kind of branching mechanism is sometimes called critical binary branching.

Then, denote by

$$X_t^{(N,\gamma)} := \frac{1}{N} \sum_i \delta_{x_t(i)} \in \mathcal{M}_F(\mathbb{R}^d), \quad t \geq 0,$$

the empirical measure of the particle system, where the sum is over all particles i alive at time t and their spatial positions at time t are given by the $x_t(i) \in \mathbb{R}^d$. Note that we denote by $\mathcal{M}_F(\mathbb{R}^d)$ resp. $\mathcal{M}_1(\mathbb{R}^d)$ the space of finite resp. probability measures on $\mathbb{R}^d$, equipped with the weak topology.

Definition 2.1. The measure-valued empirical process $\{X_t^{(N,\gamma)}, t \geq 0\}$ is called *branching Brownian motion* starting with $N \in \mathbb{N}$ particles and branching rate $\gamma > 0$.

It is well known that the corresponding *total mass process* $\{x_t^{(N,\gamma)}, t \geq 0\}$, given by

$$x_t^{(N,\gamma)} := \langle \mathbf{1}, X_t^{(N,\gamma)} \rangle, \quad t \geq 0$$

converges to a Feller diffusion if the branching rate is sped up by a factor of N, i.e., $\{x_t^{(N,N\gamma)}\} \Rightarrow \{x_t\}$, where $\{x_t\}$ solves the sde

$$dx_t = \sqrt{\gamma x_t} dB_t, \quad t \geq 0,$$

with $\{B_t\}$ a Brownian motion. Further, Watanabe showed that under a similar rescaling also the measure-valued *branching Brownian motion* converges, weakly on the space of càdlàg-paths $D_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d))$ to a non-trivial limit:

$$\{X_t^{(N,N\gamma)}\} \Rightarrow \{X_t\} \quad \text{weakly on } D_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d)).$$

Definition 2.2. The measure-valued process $\{X_t\}$ is called classical *Dawson-Watanabe process* resp. classical *super-Brownian motion.*

As limit of branching particle systems, it is not surprising that super-Brownian motion retains the branching property and is therefore infinitely divisible. In particular, if $X_0 = \hat{\mu} \in \mathcal{M}_F(\mathbb{R}^d)$, for the Laplace transform, we have

$$\mathbb{E}_{\hat{\mu}}\left[e^{-\langle \phi, X_t\rangle}\right] = e^{-\int_{\mathbb{R}^d} u_t^{\phi}\, d\hat{\mu}}, \quad t \geq 0, \tag{2.1}$$

for bounded positive measurable test functions ϕ, where $(u_t) = (u_t^{\phi})$ solves the Cauchy initial value problem

$$\begin{cases} \partial u_t = \Delta u_t - \frac{\gamma}{2} u_t^2, & t \geq 0, \\ u_0 = \phi. \end{cases} \tag{2.2}$$

An important feature is the *universality* of this limit in the following sense: Super-Brownian motion arises as limit of branching particle systems with underlying Brownian motion for *all critical branching mechanisms whose variance converges to a finite limit.* It is therefore a suitable model for large populations, if the variance of the reproductive mechanism is known and finite. In this case the total number of offspring generated by a single particle, for large populations, is small when compared with the total population size. Certainly, this property is shared with the universality of the Feller diffusion as the limit of finite-variance Bienaymé-Galton-Watson processes. For more details and further information on super-Brownian motion, see, e.g., [10] or [15].

2.2. The Fleming-Viot process

In 1979, Fleming and Viot [21] introduced their now well-known probability-measure-valued stochastic process as a model for the distribution of allelic frequencies in a selectively neutral genetic population with mutation. In their scenario, the underlying motion is interpreted as mutation in some genetic type space. In this subsection, however, we will consider only Brownian motion in $\mathbb{R}^d$ as underlying motion process.

Again, we identify the measure-valued process as limit of approximating particle systems. One of the most striking differences to super-Brownian motion is that the total mass of the Fleming-Viot process (and the approximating empirical measures) always equals one. Again, consider a system of N particles, whose initial positions are i.i.d. distributed in $\mathbb{R}^d$ according to some probability measure μ. As for branching Brownian motion, each particle is assumed to independently undergo a standard Brownian motion and is equipped with an exponential clock with rate γ. This time, however, when a clock rings, the corresponding particle chooses uniformly one of the current positions of the particles alive (including himself) and then *jumps there.* One may interpret this as a reproduction event in which the "destination particle" produces one offspring and the jumping original particle dies. Certainly, with this mechanism, the total number of particles always remains constant. Denote by

$$Y_t^{(N,\gamma)} := \frac{1}{N}\sum_{i=1}^{N} \delta_{y_t(i)} \in \mathcal{M}_1(\mathbb{R}^d), \quad t \geq 0,$$

the empirical measure of the above particle system, where the sum is over all particles i alive at time t (this time, always N-many) and their spatial positions at time t are given by the $y_t(i)$.

Definition 2.3. The measure-valued empirical process $\{Y_t^{(N,\gamma)}, t \geq 0\}$ will be called continuous-time *Moran model* starting with $N \in \mathbb{N}$ particles and branching rate $\gamma > 0$.

Similar to the super-Brownian motion case, it can be shown that the measure-valued Moran model converges weakly on the space of càdlàg-paths $D_{[0,\infty)}(M_1(\mathbb{R}^d))$ to a non-trivial limit:

$$\{Y_t^{(N,N\gamma)}\} \Rightarrow \{Y_t\} \quad \text{weakly on } D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d)).$$

Definition 2.4. The measure-valued process $\{Y_t\}$ is called *classical Fleming-Viot process.*

This time, however, we have no simple Laplace-transform characterization of the limiting process $\{Y_t\}$, since this would imply the branching property which is not fulfilled for the classical Fleming-Viot process – in particular, the total population size is constant. Hence, one usually works with a generator-based characterization:

Proposition 2.5. *The* (*classical*) *Fleming-Viot process is the Markov process* $\{Y_t\}$, *with values in* $\mathcal{M}_1(\mathbb{R}^d)$, *such that for functions* F *of the form*

$$F(\rho) := \prod_{i=1}^{n} \langle \phi_i, \rho \rangle, \quad \rho \in \mathcal{M}_1(\mathbb{R}^d), \tag{2.3}$$

where $n \in \mathbb{N}$ *and* $\phi_i \in C_c^2(\mathbb{R}^d)$, *the generator* $L^{\delta_0,\Delta}$ *of* $\{Y_t, t \geq 0\}$ *can be written as*

$$L^{\delta_0,\Delta}F(\rho) = \sum_{i=1}^{n} \langle \Delta\phi_i, \rho \rangle \prod_{j\neq i} \langle \phi_j, \rho \rangle + \sum_{1\leq i<j\leq n} \Big[\langle \phi_i\phi_j, \rho \rangle - \langle \phi_i, \rho \rangle \langle \phi_j, \rho \rangle \Big] \prod_{k\neq i,j} \langle \phi_k, \rho \rangle,$$

with Δ *the Laplace operator.*

Note that the rôle of the superscript "δ_0" will become clear once we identify this process as a special case of a much larger class of (generalized) Fleming-Viot processes below.

Again, the Fleming-Viot process is a universal limit which arises from approximating particle systems of constant population size. Also in this case the convergence of the variance of the reproduction events is crucial. It should be noted that not only continuous-time particle systems converge, but also suitably rescaled discrete-time systems, such as the famous Wright-Fisher model (see, e.g., [42]) and many other Cannings' models [8], [9]. For a classification of the possible limits of population models, see [29].

3. Genealogies and exchangeable coalescent processes

For population models of fixed population size in the domain of attraction of the classical Fleming-Viot process, such as the Moran model described above, the family tree, or *genealogy*, of a finite sample taken from the population at time t can be described by the now classical *Kingman-coalescent*, which we introduce briefly, followed by the more recently discovered and much more general so-called *Lambda-* and *Xi-coalescents.*

3.1. Kingman's coalescent

Let $\mathcal{P}_n$ be the set of partitions of $\{1,\dots,n\}$ and let $\mathcal{P}$ denote the set of partitions of $\mathbb{N}$. For each $n \in \mathbb{N}$, Kingman [23] introduced the so-called *n-coalescent*, which is a $\mathcal{P}_n$-valued continuous time Markov process $\{\Pi_n(t), t \geq 0\}$, such that $\Pi_n(0)$ is the partition of $\{1,\dots,n\}$ into singleton blocks, and then each pair of blocks merges at rate one. Given there are b blocks at present, this means that the overall rate to see a merger between blocks is $\binom{b}{2}$. Note that only *binary mergers* are allowed. Kingman [23] also showed that there exists a corresponding $\mathcal{P}$-valued Markov process:

Definition 3.1. The projective limit $\{\Pi(t)\}$ on $\mathcal{P}$ of the n-coalescents $\{\Pi_n(t)\}$ is called (standard) *Kingman-coalescent.*

To see that this definition makes sense, note that the restriction of any n-coalescent to $\{1,\dots,m\}$, where $1 \leq m \leq n$, is an m-coalescent. Hence the process can be constructed by an application of the standard extension theorem.

It is well known (cf. [11]) that the classical Fleming-Viot process is dual to *Kingman's coalescent*, introduced in [23], in the following (informal) sense: For $t \geq 0$, if one takes a uniform sample of n individuals from Y_t and forgets about the respective spatial positions of the n particles, then their genealogical tree backwards in time can be viewed as a realization of *Kingman's n-coalescent.* That means, at each time $t-s$, where $s \in [0,t]$ (hence *backwards* in time), the ancestral lineages of each particle merge at infinitesimal rate $\binom{k}{2}$, where $k \in \{2,\dots,n\}$ denotes the number of distinct lineages present at time $t-s(-)$. This can be made rigorous, for example, using Donnelly and Kurtz' *lookdown construction* [12], as we shall explain later in the case of so-called Lambda-coalescents, and spatial information may also be incorporated, see, e.g., [15], Section 1.12.

3.2. Lambda-coalescents

Pitman [32] and Sagitov [38] introduced and discussed exchangeable coalescents which allow *multiple mergers*, i.e., more than just two blocks may merge at a time. Again, a coalescent with multiple mergers is a $\mathcal{P}$-valued Markov-process $\{\Pi^\Lambda(t), t \geq 0\}$, such that for each n, its restriction to the first n positive integers is a $\mathcal{P}_n$-valued Markov process with the following transition rates: Whenever there are b blocks in the partition at present, each k-tuple of blocks (where $2 \leq k \leq b \leq n$) is merging to form a single block at rate $\lambda_{b,k}$, and no other transitions are possible. The rates $\lambda_{b,k}$ do neither depend on n nor on the structure of the blocks. Pitman showed that in order to be consistent when switching from a subsample of size $b+1$ to a sub-subsample of size b, that is, for all $b, k \geq 2, b \geq k$,

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1},$$

such transition rates must necessarily satisfy

$$\lambda_{b,k} = \int_0^1 x^k (1-x)^{b-k} \frac{1}{x^2} \Lambda(dx), \tag{3.1}$$

for some finite measure Λ on the unit interval. Note that (3.1) sets up a one-to-one correspondence between coalescents with multiple collisions and finite measures Λ. Indeed, it is easy to see that the $\lambda_{b,k}$ determine Λ since they satisfy the conditions of Hausdorff's moment problem, which has a unique solution.

Definition 3.2. For each $\Lambda \in \mathcal{M}_F([0,1])$, the associated coalescent process $\{\Pi^\Lambda(t)\}$ is called *Lambda-coalescent* (resp. Λ-*coalescent*).

Note that the family of Lambda-coalescents is rather large, and in particular cannot be parametrised by a few real variables. Important examples include $\Lambda = \delta_0$ (Kingman's coalescent) and $\Lambda = \delta_1$ (leading to star-shaped genealogies, i.e., one huge merger into one single block). Later, we will be concerned with an important parametric subclasses of Λ-coalescents, namely the so-called *Beta-coalescents*, where Λ has a Beta$(2-\beta, \beta)$-density for some $\beta \in (0,2]$. To avoid trivialities, we will henceforth assume that $\Lambda \neq 0$.

Remark 3.3. An important difference between the classical Kingman-coalescent and coalescents which allow for multiple mergers should be pointed out here. Roughly speaking, a Kingman coalescent arises as limiting genealogy from a so-called Cannings population model ([8], [9]), if the individuals produce a number of offspring that is negligible when compared to the total population size (in particular, if the variance of the reproduction mechanism converges to a finite limit). More general coalescents with multiple mergers arise, once the offspring distribution is such that a non-negligible proportion, say $x \in (0,1]$, of the population alive in the next generation goes back to a single reproduction event from a single ancestor. In this case, $x^{-2}\Lambda(dx)$ can be interpreted as the intensity at which we see such proportions x. Precise conditions on the underlying Cannings-models and the classification of the corresponding limiting genealogies can be found in [29].

Definition 3.4. We say that a Lambda-coalescent $\{\Pi^\Lambda(t)\}$ "comes down from infinity", if for any initial partition $\Pi^\Lambda(0)$ in $\mathcal{P}$, and for all $\varepsilon > 0$, the partition $\Pi^\Lambda(\varepsilon)$ a.s. consists of finitely many blocks only.

Lemma 3.5 (Schweinsberg [39]**).** *Given $\Lambda \in \mathcal{M}_F([0,1])$, the corresponding Λ-coalescent comes down from infinity iff*

- *either Λ has an atom at 0, or*
- *Λ has no atom at zero and*

$$\sum_{b=2}^{\infty}\left[\sum_{k=2}^{b}(k-1)\binom{b}{k}\int_{[0,1]} x^{k-2}(1-x)^{b-k}\Lambda(dx)\right]^{-1} < \infty. \tag{3.2}$$

Note that if a coalescent comes down from infinity, the time to come down to only one block has finite expectation.

Examples for coalescents which satisfy (3.2) are Kingman's coalescent and the so-called Beta$(2-\beta,\beta)$-coalescents with $\beta \in (1,2]$, where

$$\Lambda(dx) = \frac{\Gamma(2)}{\Gamma(2-\beta)\Gamma(\beta)} x^{1-\beta}(1-x)^{\beta-1}\,dx. \tag{3.3}$$

An important example for a coalescent, which (only just) does not come down from infinity is the so-called Bolthausen-Sznitman coalescent, where $\Lambda(dx) = dx$ on $[0,1]$. In the above terminology, it is the Beta-coalescent with $\beta = 1$.

3.3. Xi-coalescents

The largest class of exchangeable coalescents are given by the so-called Xi-coalescents (resp. Ξ-coalescents), introduced in [40] and [29]. Extending the Λ-coalescents, these coalescents do not only include situations in which several lineages coalesce to one lineage at a time, but where even simultaneous multiple collisions are allowed.

Formally, Schweinsberg [40] showed that any $\mathcal{P}$-valued exchangeable coalescent $\{\Pi_t^\Xi\}$ is characterized by a finite measure Ξ on the infinite simplex

$$\Delta \;:=\; \big\{\zeta = (\zeta_1, \zeta_2, \ldots) : \zeta_1 \ge \zeta_2 \ge \cdots \ge 0, \textstyle\sum_{i=1}^{\infty}\zeta_i \le 1\big\}. \tag{3.4}$$

Throughout the paper, for $\zeta \in \Delta$, the notation $|\zeta| := \sum_{i=1}^{\infty}\zeta_i$ and $(\zeta,\zeta) := \sum_{i=1}^{\infty}\zeta_i^2$ will be used for convenience.

Coalescent processes with multiple collisions (Λ-coalescents) occur if the measure Ξ is concentrated on the subset of all points $\zeta \in \Delta$ satisfying $\zeta_i = 0$ for all $i \ge 2$. The Kingman-coalescent corresponds to the case $\Xi = \delta_{\mathbf{0}}$. It is convenient to decompose the measure Ξ into a 'Kingman part' and a 'simultaneous multiple collision part', that is,

$$\Xi = a\delta_{\mathbf{0}} + \Xi_0 \qquad \text{with} \qquad a := \Xi(\{\mathbf{0}\}) \qquad \text{and} \qquad \Xi_0(\{\mathbf{0}\}) = 0. \tag{3.5}$$

The transition rates of the Ξ-coalescent Π^Ξ are given as follows. Suppose there are currently b blocks. Exactly $\sum_{i=1}^{r} k_i$ blocks collide into r new blocks, each containing $k_1, \ldots, k_r \ge 2$ original blocks, and s single blocks remain unchanged, such that the condition $\sum_{i=1}^{r} k_i + s = b$ holds. The order of $k_1, \ldots, k_r$ does not matter.

The rate at which the above collision happens is then given as (Schweinsberg [40, Theorem 2]):

$$\lambda_{b;k_1,\dots,k_r;s} = a\mathbf{1}_{\{r=1,k_1=2\}} + \int_\Delta \sum_{l=0}^{s} \binom{s}{l} (1-|\zeta|)^{s-l} \sum_{i_1 \neq \cdots \neq i_{r+l}} \zeta_{i_1}^{k_1} \cdots \zeta_{i_r}^{k_r} \zeta_{i_{r+1}} \cdots \zeta_{i_{r+l}} \frac{\Xi_0(d\zeta)}{(\zeta,\zeta)}. \tag{3.6}$$

Genealogies with simultaneous multiple mergers arise, for example, in populations undergoing short, but severe bottlenecks: in this case, only “a few” particles (compared to the original population size) survive during the bottleneck. See [6] for details.

4. More general measure-valued processes

4.1. Generalized Fleming-Viot processes with self-similar motion

Since its introduction, the classical Fleming-Viot process received a great deal of attention from both geneticists and probabilists. As already pointed out, one reason is that it is the natural limit of a large class of exchangeable population models with constant size and finite-variance reproduction mechanism, in particular the so-called Moran-model.

More general limit population processes describing situations where a single individual may produce a non-negligible fraction of the total population have been introduced in [13] (see also [1] for a different approach). We call these processes *Lambda-Fleming-Viot processes* (resp. Λ*-Fleming-Viot processes*). The limits of the corresponding dual genealogical processes have been classified in [38] and [29] – see [3] for an overview.

Definition 4.1. Let $\Lambda \in \mathcal{M}_F([0,1])$. The corresponding Lambda-Fleming-Viot process is a probability measure-valued Markov process $\{Y_t^{\Lambda,\Delta_\alpha}\}$ with paths in $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$, whose generator $L^{\Lambda,\Delta_\alpha}$ acts on functions F of the form (2.3) as

$$L^{\Lambda,\Delta_\alpha} F(\rho) = \sum_{i=1}^{n} \langle \Delta_\alpha \phi_i, \rho\rangle \prod_{j\neq i} \langle \phi_j, \rho\rangle + \sum_{\substack{J\subset\{1,\dots,n\}\\ |J|\geq 2}} \lambda_{n,|J|} \Big[\langle \prod_{j\in J} \phi_j, \rho\rangle - \prod_{j\in J} \langle \phi_j, \rho\rangle \Big] \prod_{k\notin J} \langle \phi_k, \rho\rangle, \tag{4.1}$$

where

$$\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\Lambda(dx), \quad n \geq k \geq 2, \tag{4.2}$$

with Λ a finite measure on $[0,1]$, and where $\Delta_\alpha = -(-\Delta)^{\alpha/2}$ is the fractional Laplacian of index $\alpha \in (0,2]$.

Note that in this definition, we also allow more general underlying motion processes. In fact, for $\alpha \in (0,2]$, the fractional Laplacian of index α is the generator of a standard symmetric α-stable process in $\mathbb{R}^d$, see, e.g., [44], Chapter IX.11, or [18], Chapter IX.6.

Proposition 4.2. *For $\alpha \in (0,2]$ the standard symmetric α-stable process $\{B_t^{(\alpha)}\}$ is statistically self-similar, that is, for $k > 0$, its law satisfies the scaling property*

$$\mathcal{L}\bigg(\frac{B_{kt}^{(\alpha)}}{k^{1/\alpha}}, t \geq 0\bigg) = \mathcal{L}\big(B_t^{(\alpha)}, t \geq 0\big). \tag{4.3}$$

For given $\Lambda \in \mathcal{M}_F([0,1])$, note that the rates $\lambda_{n,k}$ from Definition 4.1 precisely describe the transitions of $\{\Pi_t^\Lambda, t \geq 0\}$, the Λ-*coalescent* from Definition 3.2. Hence, a Λ-Fleming-Viot process is dual to a Λ-coalescent (as shown in [13], pp. 195 and [1]), similar to the duality between the classical Fleming-Viot process and Kingman's coalescent established in [11]. Note that Kingman's coalescent corresponds to the choice $\Lambda = \delta_0$, and we recover (up to the underlying motion) the generator given in Proposition 2.5. In the notation of Definition 4.1, our classical Fleming-Viot process is denoted by $\{Y_t\} = \{Y_t^{\delta_0,\Delta}\}$.

4.2. The Beta-Fleming-Viot process

A particularly important subclass of Fleming-Viot processes is given by the so-called Beta$(2-\beta,\beta)$-Fleming-Viot processes:

Definition 4.3. Assume Λ is a probability measure on $[0,1]$ with density

$$\Lambda(dx) = \frac{\Gamma(2)}{\Gamma(2-\beta)\Gamma(\beta)} x^{1-\beta}(1-x)^{\beta-1}\,dx$$

for some $\beta \in (0,2]$. The corresponding coalescent process is called *Beta*$(2-\beta,\beta)$-*coalescent*, the corresponding dual measure-valued process is called *Beta*$(2-\beta,\beta)$-*Fleming-Viot process* or just Beta-Fleming-Viot process if β is known.

4.3. Xi-Fleming-Viot processes

In much the same way as Lambda-coalescents are dual to Lambda-Fleming-Viot processes, Xi-coalescents are dual to Xi-Fleming-Viot processes. The latter have been constructed explicitly in [6] and may be described by their generator as follows:

Definition 4.4. Let $\Xi = a\delta_0 + \Xi_0$ be a finite measure on the simplex defined in (3.4) and (3.5). The corresponding Xi-Fleming-Viot process is a probability measure-valued Markov processes $\{Y_t^{\Xi,\Delta_\alpha}\}$ with paths in $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$, whose generator L^{Ξ,Δ_α} acts on functions F of the form (2.3) as

$$L^{\Xi,\Delta_\alpha}F(\rho) \;:=\; L^{a\delta_0}F(\rho) + L^{\Xi_0}F(\rho) + L^{\Delta_\alpha}F(\rho),$$

where

$$L^{a\delta_0}F(\rho) := \sum_{1\le i<j\le n}\Big[\langle\phi_i\phi_j,\rho\rangle - \langle\phi_i,\rho\rangle\langle\phi_j,\rho\rangle\Big]\prod_{k\neq i,j}\langle\phi_k,\rho\rangle,$$

$$L^{\Xi_0}F(\rho) := \int_\Delta\int_{(\mathbb{R}^d)^{\mathbb{N}}}\left[F\big((1-|\zeta|)\rho+\textstyle\sum_{i=1}^\infty\zeta_i\delta_{x_i}\big)-F(\rho)\right]\rho^{\otimes\mathbb{N}}(d\mathbf{x})\dfrac{\Xi_0(d\zeta)}{(\zeta,\zeta)},$$

and finally

$$L^{\Delta_\alpha}F(\rho) := \sum_{i=1}^n\langle\Delta_\alpha\phi_i,\rho\rangle\prod_{j\neq i}\langle\phi_j,\rho\rangle,$$

where, again, $\Delta_\alpha = -(-\Delta)^{\alpha/2}$ is the fractional Laplacian of index $\alpha\in(0,2]$.

4.4. Infinitely-divisible superprocesses

Let us come back to superprocesses, i.e., measure-valued processes which have the branching property. So far, we have only dealt with super-Brownian motion, where the underlying motion is Brownian and the branching mechanism, recalling (2.1) and (2.2), has Laplace exponent $\frac{u^2}{2}$ (a so-called stable branching mechanism of index 2). As for the Fleming-Viot processes, we wish to extend this to a more general framework, in particular allowing stable underlying motions and more general branching mechanisms of infinite variance.

Definition 4.5. For $X_0^{\beta,\Delta_\alpha} := \hat\mu \in \mathcal{M}_F(\mathbb{R}^d)$ and $\beta\in(1,2]$ denote by $\{X_t^{\beta,\Delta_\alpha}\}$ the so-called *(α,d,β)-superprocess* characterized by the Laplace transform

$$\mathbb{E}_{\hat\mu}\left[e^{-\langle\phi,X_t^{\beta,\Delta_\alpha}\rangle}\right] = e^{-\int_{\mathbb{R}^d}u_t^\phi\,d\hat\mu}, \tag{4.4}$$

for bounded positive measurable test functions ϕ, where $(u_t)=(u_t^\phi)$ solves the Cauchy initial value problem

$$\begin{cases}\partial u_t = \Delta_\alpha u_t - \frac{1}{\beta}(u_t)^\beta, & t\ge 0,\\ u_0=\phi.\end{cases} \tag{4.5}$$

For a construction of such superprocesses and particle approximations, see, e.g., [10]. In particular, the (α,d,β)-superprocess can be obtained from a branching particle system in a similar way as in Section 2.1, under suitable rescaling, where the branching mechanism has offspring generating function

$$\Phi(s) = \frac{1}{\beta}(1-s)^\beta + s, \quad s\ge 0,\ \beta\in(1,2].$$

See, e.g., [15] for details. In particular, for $\beta\in(1,2)$, the offspring distribution has infinite variance.

Fleischmann and Wachtel [20] used super-critical (α,d,β)-superprocesses to obtain by approximation, letting $\beta\downarrow 1$, a superprocesses with *Neveu's branching mechanism*:

Definition 4.6 (Neveu superprocess). For $X_0^{1,\Delta_\alpha} := \hat{\mu} \in \mathcal{M}_F(\mathbb{R}^d)$, the Neveu Superprocess $\{X_t^{1,\Delta_\alpha}\}$ has Laplace transform

$$\mathbb{E}_{\hat{\mu}}\Big[e^{-\langle\phi,X_t^{1,\Delta_\alpha}\rangle}\Big] = e^{-\int_{\mathbb{R}^d} u_t^\phi \, d\hat{\mu}},$$

for bounded positive measurable test functions ϕ, where $(u_t) = (u_t^\phi)$ solves the Cauchy initial value problem

$$\begin{cases} \partial u_t = \Delta_\alpha u_t - u_t \log u_t, & t \geq 0, \\ u_0 = \phi. \end{cases}$$

The (α, d, β)-superprocesses for $\alpha \in (0,2], \beta \in [1,2]$ will be the most general class of superprocesses that we will be concerned with in this article, since it has interesting relations with the Beta-Fleming-Viot processes, as we will see below. For superprocesses with general branching mechanisms (usually denoted by Ψ), see, e.g., [10].

4.5. The relation between the two classes of measure-valued process

The close relation between the approximating particle systems presented in Section 2 suggests a close connection of the limiting processes. This is not entirely obvious from the generator resp. Laplace transform characterizations. However, for the classical measure-valued processes, this connection has already been clarified in the early nineties. To simplify the exposition, we will for a moment suppress the underlying motion processes of our superprocesses.

Let $\{Y_t\} = \{Y_t^{\delta_0,*}\}$ be the classical Fleming-Viot process, where the underlying motion of the particles is constant, i.e., the measure-valued process with generator

$$L^{*,\Delta}F(\rho) = \sum_{1\leq i<j\leq n} \Big[\langle\phi_i\phi_j,\rho\rangle - \langle\phi_i,\rho\rangle\langle\phi_j,\rho\rangle\Big] \prod_{k\neq i,j} \langle\phi_k,\rho\rangle,$$

and similarly, let $\{X_t\} = \{X_t^{2,*}\}$ be the classical Dawson-Watanabe process with constant underlying motion.

Consider the *ratio process* $\frac{X_t}{\langle\mathbf{1},X_t\rangle}$, which is well defined as long as $\langle\mathbf{1},X_t\rangle > 0$. Define a time change by

$$T_t := \int_0^t \frac{1}{\langle\mathbf{1},X_s\rangle}\, ds, \quad T^{-1}(t) := \inf\{s : T_s > t\}.$$

Generalizing a result of [16], Perkins [31] proved that, conditioned on the total population size process $\{\langle\mathbf{1},X_t\rangle, t \geq 0\}$, the time-changed ratio process agrees in law with the classical Fleming-Viot process, i.e.,

$$\left\{\frac{X_{T^{-1}(t)}}{\langle\mathbf{1},X_{T^{-1}}(t)\rangle}, t \geq 0\right\} \overset{d}{=} \{Y_t, t \geq 0\}.$$

If we would consider underlying motions, then this result would give equality in law with a Fleming-Viot process with time-inhomogeneous underlying Brownian motion.

It is a natural question to ask whether such a close relationship also holds for more general Fleming-Viot and superprocesses. It has been treated in the works [13], [22] and [5]. In [5], a complete answer is given:

Theorem 4.1. *The ratio process of a general superprocess can be time-changed with an additive functional of its total mass process to obtain a Markov process, iff the branching mechanism is continuous β-stable.*

In particular, for $\beta \in [1,2]$, if $\{X_t^{\beta,*}\}$ is the superprocess with β-stable branching (with constant underlying motion) and $\{Y_t^{\beta,*}\}$ is the Beta$(2-\beta,\beta)$-Fleming-Viot process (with constant underlying motion), then, defining

$$T_t := \int_0^t \langle \mathbf{1}, X_s^{\beta,*}\rangle^{1-\beta}\, ds, \quad T^{-1}(t) := \inf\{s : T_s > t\}, \tag{4.6}$$

we have

$$\left\{ \frac{X_{T^{-1}(t)}^{\beta,*}}{\langle \mathbf{1}, X_{T^{-1}}^{\beta,*}(t)\rangle}, t \geq 0 \right\} \stackrel{d}{=} \{Y_t^{\beta,*}, t \geq 0\}.$$

When there are underlying motions involved, the result yields equality in law with a Fleming-Viot process with time-inhomogeneous underlying motions. However, note that for $\beta = 1$, *the time-change becomes trivial.* Hence, we obtain

Proposition 4.7. *The normalized Neveu superprocess is a Beta$(1,1)$-Fleming-Viot process, i.e., for $\alpha \in (0,2]$,*

$$\left\{ \frac{X_t^{1,\Delta_\alpha}}{\langle 1, X_t^{1,\Delta_\alpha}\rangle}, t \geq 0 \right\} \stackrel{d}{=} \{Y_t^{1,\Delta_\alpha}, t \geq 0\}.$$

Note that this implies the well-known fact that the critical Neveu branching process does not die out in finite time.

5. Donnelly and Kurtz' lookdown construction

5.1. A countable representation for the Λ-Fleming-Viot process

Donnelly and Kurtz' lookdown construction exploits the exchangeability of the approximating particle systems of our measure-valued processes in order to construct the limiting object a.s. via de Finetti's theorem. To this end, one needs to construct the approximating empirical measure on the same probability space.

We consider a countably infinite system of individuals, each particle being identified by a level $j \in \mathbb{N}$. We equip the levels with types ξ_t^j in $\mathbb{R}^d$, $j \in \mathbb{N}$. Initially, we require the types $\xi_0 = (\xi_0^j)_{j\in\mathbb{N}}$ to be an i.i.d. vector (in particular exchangeable), so that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^N \delta_{\xi_0^j} = \mu,$$

for some finite measure $\mu \in \mathcal{M}_1(\mathbb{R}^d)$, which will be the initial condition of the generalized Fleming-Viot process constructed below via (5.6). The point is that the construction will preserve exchangeability.

There are two "sets of ingredients" for the reproduction mechanism of these particles, one corresponding to the "finite variance" part $\Lambda(\{0\})$, and the other to the "extreme reproductive events" described by $\Lambda_0 = \Lambda - \Lambda(\{0\})\delta_0$. Restricted to the first N levels, the dynamics is that of a very particular permutation of a generalized Moran model with the property that the particle with the highest level is always the next to die.

For the first part, let $\{L_{ij}(t), t \geq 0\}$, $1 \leq i < j < \infty$, be independent Poisson processes with rate $\Lambda(\{0\})$. Intuitively, at jump times of L_{ij}, the particle at level j "looks down" to level i and copies the type from there, corresponding to a single birth event in a(n approximating) Moran model. Let $\Delta L_{ij}(t) = L_{ij}(t) - L_{ij}(t-)$. At jump times, types on levels above j are shifted accordingly, in formulas

$$\xi_t^k = \begin{cases} \xi_{t-}^k, & \text{if } k < j, \\ \xi_{t-}^i, & \text{if } k = j, \\ \xi_{t-}^{k-1}, & \text{if } k > j, \end{cases} \tag{5.1}$$

if $\Delta L_{ij}(t) = 1$. This mechanism is well defined because for each k, there are only finitely many processes L_{ij}, $i < j \leq k$ at whose jump times ξ^k has to be modified.

For the second part, which corresponds to multiple birth events, let n be a Poisson point process on $\mathbb{R}^+ \times (0,1] \times [0,1]^{\mathbb{N}}$ with intensity measure $dt \otimes r^{-2}\Lambda_0(dr) \otimes (du)^{\otimes \mathbb{N}}$. Note that for almost all realizations $\{(t_i, y_i, (u_{ij}))\}$ of n, we have

$$\sum_{i\,:\,t_i \leq t} y_i^2 < \infty \quad \text{for all } t \geq 0. \tag{5.2}$$

The jump times t_i in our point configuration n correspond to reproduction events. Define for $J \subset \{1, \dots, l\}$ with $|J| \geq 2$,

$$L_J^l(t) := \sum_{i\,:\,t_i \leq t} \prod_{j \in J} 1_{\{u_{ij} \leq y_i\}} \prod_{j \in \{1,\dots,l\} - J} 1_{\{u_{ij} > y_i\}}. \tag{5.3}$$

$L_J^l(t)$ counts how many times, among the levels in $\{1, \dots, l\}$, exactly those in J were involved in a birth event up to time t. Note that for any configuration n satisfying (5.2), since $|J| \geq 2$, we have

$$\mathbb{E}\big[L_J^l(t) \,\big|\, n|_{[0,t] \times (0,1] \times [0,1]^{\mathbb{N}}}\big] = \sum_{i\,:\,t_i \leq t} y_i^{|J|}(1 - y_i)^{l-|J|} \leq \sum_{i\,:\,t_i \leq t} y_i^2 < \infty,$$

so that $L_J^l(t)$ is a.s. finite.

Intuitively, at a jump t_i, each level performs a "uniform coin toss", and all the levels j with $u_{ij} \leq y_i$ participate in this birth event. Each participating level adopts the type of the smallest level involved. All the other individuals are shifted upwards accordingly, keeping their original order with respect to their levels (see Figure 1). More formally, if $t = t_i$ is a jump time and j is the smallest level

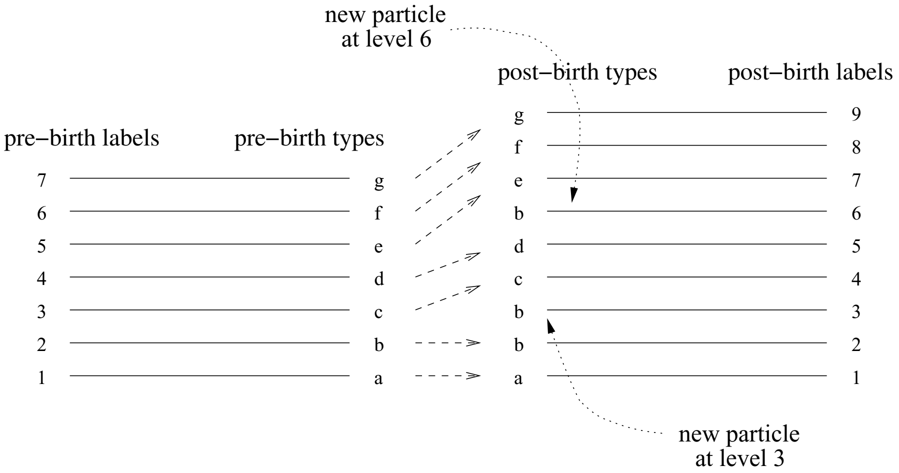

FIGURE 1. Relabeling after a birth event involving levels 2, 3 and 6.

involved, i.e., $u_{ij} \le y_i$ and $u_{ik} > y_i$ for $k < j$, we put

$$\xi_t^k = \begin{cases} \xi_{t-}^k, & \text{for } k \le j, \\ \xi_{t-}^j, & \text{for } k > j \text{ with } u_{ik} \le y_i, \\ \xi_{t-}^{k-J_t^k}, & \text{otherwise}, \end{cases} \tag{5.4}$$

where $J_{t_i}^k = \#\{m < k : u_{im} \le y_i\} - 1$. Let us define $\mathcal{G} = (\mathcal{G}_{u,v})_{u<v}$, where for $u < v$

$$\begin{aligned} \mathcal{G}_{u,v} =& \sigma\big\{L_{ij}(t) - L_{ij}(s), u < s \le t \le v, i, j \in \mathbb{N}\big\} \\ & \vee \sigma\big\{n([s,t) \times A \times B), u < s \le t \le v, A \subset (0,1], B \subset [0,1]^{\mathbb{N}}\big\} \end{aligned} \tag{5.5}$$

is the σ-algebra describing all “genealogical events” between times u and v.

So far, we have only treated the reproductive mechanism of the particle system. In between reproduction events, all the levels follow independent α-stable motions. For a rigorous formulation, all three mechanisms together can be cast into a suitable countable system of stochastic differential equations driven by Poisson processes and α-stable processes, see [13], Section 6.

Then, for each $t > 0$, $(\xi_t^1, \xi_t^2, \dots)$ is an exchangeable random vector and

$$Z_t = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi_t^j}, \quad t \ge 0 \tag{5.6}$$

exists almost surely on $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$, and $\{Z_t, t \ge 0\}$ is the Markov process with generator (4.1) and initial condition $Z_0 = \mu$, see [13], Theorem 3.2.

5.2. Pathwise embedding of Λ-coalescents in Λ-Fleming-Viot processes

Note that for each $t > 0$ and $s \leq t$, the modified lookdown construction encodes the ancestral partition of the levels at time t with respect to the ancestors at time s before t via

$$N_i^t(s) = \text{level of level } i\text{'s ancestor at time } t - s.$$

For fixed t, the vector-valued process $\{N_i^t(s) : i \in \mathbb{N}\}_{0 \leq s \leq t}$ satisfies an "obvious" system of Poisson-process driven stochastic differential equations, see [13], p. 195, (note that we have indulged in a time re-parametrisation), and the partition-valued process defined by

$$\big\{\{i : N_i^t(s) = j\}, j = 1, 2, \dots\big\}_{0 \leq s \leq t} \tag{5.7}$$

is a standard Λ-coalescent with time interval $[0, t]$. This implies in particular by Kingman's theory of exchangeable partitions (see [23], or, e.g., [34] for an introduction), that the empirical family sizes

$$A_j^t(s) := \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} 1_{\{N_i^t(s) = j\}} \tag{5.8}$$

exist a.s. in $[0, 1]$ for each j and $s \leq t$, describing the relative frequency at time t of descendants of the particle at level j at time $t - s$.

5.3. Lookdown-constructions for the classical- and the Ξ-Fleming-Viot process

In the preceding section, we provided a brief description of the (modified) lookdown construction for Lambda-Fleming-Viot processes. In an earlier paper [12], Donnelly and Kurtz already provided an (unmodified) lookdown construction for the classical Fleming-Viot process. The above construction can be extended to far more general measure-valued processes, even including interaction between particles. See [15] for an introduction. Finally, a recent detailed extension of this method to Xi-Fleming-Viot processes has been given in [6]

6. Application: Measure-valued processes as wandering and flickering random measures

6.1. Dawson and Hochberg's wandering random measures

Dawson and Hochberg [11] demonstrated that the classical Fleming-Viot process performs a "wandering", but "coherent" motion that, appropriately rescaled, approaches Brownian motion.

Definition 6.1 (Coherent wandering random measure). Let $\{Z_t\}$ be a measure-valued process on $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$. The process is called a *coherent wandering random measure*, if there is a "centering process" $\{z(t), t \geq 0\}$ with values in $\mathbb{R}^d$ and for each $\varepsilon > 0$ a real-valued stationary "radius process" $\{R_\varepsilon(t), t \geq 0\}$ and an a.s. finite T_0, such that

$$Z_t\big(B_{z(t)}(R_\varepsilon(t))\big) \geq 1 - \varepsilon \quad \text{for } t \geq T_0 \quad \text{a.s.}, \tag{6.1}$$

where $B_x(r)$ is the closed ball of radius r around $x \in \mathbb{R}^d$. Moreover, if one may choose $\varepsilon = 0$ in (6.1), the process is called *compactly coherent.* If the process is not coherent, it is called *dissipative.*

One natural choice for $\{z(t), t \geq 0\}$ is the "center of mass process" $z(t) = \int z\, Z_t(dz)$. With this process, one obtains easily the following proposition as an obvious extension of the proof of Theorem 7.2 in [11]:

Proposition 6.2. *The "semi-classical" Fleming-Viot process $\{Y_t^{\delta_0,\Delta_\alpha}\}$ is a coherent wandering random measure. Further, for $\alpha = 2$, the classical Fleming-Viot process $\{Y_t^{\delta_0,\Delta}\}$ is a compactly coherent wandering random measure.*

In the context of the lookdown construction, often a more convenient choice for the centering process is $z(t) = \xi_t^1$, the location of the so-called "level-1 particle" (see Section 5). The following simple result can be found in [4].

Proposition 6.3. *Let $\{Y_t^{\Lambda,\Delta_\alpha}\}$ be realized with the help of the lookdown construction. Then, if one chooses the position of the "level-*1*" particle as centering process, $\{Y_t^{\Lambda,\Delta_\alpha}\}$ is a coherent wandering random measure.*

If the process $\{Y_t^{\Lambda,\Delta_\alpha}\}$ has the compact support property, this will also yield *compact coherence*, i.e., one can choose $\varepsilon = 0$ in (6.1), see [11, Theorem 7.1]. We will turn to the support properties of the generalized Fleming-Viot properties in the next section.

6.2. Longterm-scaling for self-similar motion mechanisms

The long-time behaviour of a generalized Fleming-Viot process is governed by the interplay between motion and resampling mechanism. If the underlying motion is statistically self-similar, as in the case of a stable Lévy-process of index α, then one may attempt to capture this via a suitable space-time rescaling.

Definition 6.4. Let $\Lambda \in \mathcal{M}_F([0,1])$ and define the time-space rescaled process $\{Y_t^{\Lambda,\Delta_\alpha}[k], t \geq 0\}$ via

$$\langle \phi, Y_t^{\Lambda,\Delta_\alpha}[k] \rangle := \langle \phi(\cdot/k^{1/\alpha}), Y_{kt}^{\Lambda,\Delta_\alpha} \rangle, \tag{6.2}$$

for $\phi \in b\mathcal{B}(\mathbb{R}^d)$ and $t \geq 0$.

Recall that for $\alpha \in (0,2]$ we let $B^{(\alpha)} = \{B_t^{(\alpha)}\}$ be the standard symmetric stable process of index α, starting from $B_0^{(\alpha)} = 0$.

Proposition 6.5 (Scaling). *Fix $\mu \in \mathcal{M}_1(\mathbb{R}^d)$ as the initial condition of the un-scaled process $\{Y_t^{\Lambda,\Delta_\alpha}\}$. Then due to the statistical self-similarity* (4.3) *of the underlying motion process, for each k, the process $\{Y_t^{(k)}, t \geq 0\}$, defined by*

$$Y_t^{(k)} = Y_t^{k\Lambda,\Delta_\alpha}, \quad t \geq 0, \tag{6.3}$$

starting from the image measure of μ under $x \mapsto x/k^{1/\alpha}$, has the same distribution as $\{Y_t^{\Lambda,\Delta_\alpha}[k], t \geq 0\}$ defined in (6.2)*.*

It will be convenient to work in the following with a version of $Y^{(k)}$ which is obtained from a lookdown construction with "parameter" $k\Lambda$, in particular, we have

$$Y_t^{(k)} = \lim_{n\to\infty} \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_t^i}, \quad t \geq 0.$$

Note that the family ξ^i, $i \in \mathbb{N}$, used to construct $Y^{(k)}$, depends (implicitly) on k, but for the sake of readability, we suppress this in our notation.

With these definitions, one readily expects the following convergence of finite-dimensional distributions (f.d.d.) to hold, which is shown in [4].

Proposition 6.6 (Longterm-scaling). *For each finite collection of time-points* $0 \leq t_1 < \cdots < t_n$*, we have*

$$\big(Y_{t_1}^{\Lambda,\Delta_\alpha}[k], \ldots, Y_{t_n}^{\Lambda,\Delta_\alpha}[k]\big) \Rightarrow \big(\delta_{B_{t_1}^{(\alpha)}}, \ldots, \delta_{B_{t_n}^{(\alpha)}}\big) \quad \textit{as} \quad k \to \infty. \tag{6.4}$$

Note that for the classical $\{Y_t^{\delta_0,\Delta}, t \geq 0\}$, this is essentially Theorem 8.1 in [11]. Combining Proposition 4.7 and Proposition 6.5, we recover Part (a) of Theorem 1 in [20].

In addition to f.d.d.-convergence, Part (b) of of Theorem 1 in [20] provides weak convergence on $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$ if the underlying motion of the spatial Neveu process is Brownian. However, the question whether this holds in general seems to be inaccessible to the Laplace-transform and moment-based methods of [20], and therefore had been left open. Rather surprisingly, in [4] it is shown that pathwise convergence does *not* hold if $\alpha < 2$, and that, with the help of Donnelly and Kurtz' modified lookdown construction ([13]), it is possible to understand explicitly "what goes wrong" in terms of "sparks" and of a family of "flickering random measures".

To complete the picture, we cite again [4] to provide the full classification of the scaling behaviour of generalized α-stable Fleming-Viot processes.

Theorem 6.1 (Convergence on path space). *Under the above conditions,* (6.4) *holds weakly on* $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$ *if and only if* $\alpha = 2$.

6.3. Sparks and flickering random measures

The following terminology is taken from [4].

Definition 6.7 (Sparks). Consider a path $\omega = \{\omega_t, t \geq 0\}$ in $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$. We say that ω exhibits an *ε-δ-spark* (on the interval $[0,T]$) if there exist time points $0 < t_1 < t_2 < t_3 \leq T$ such that $t_3 - t_1 \leq \delta$

$$d_{\mathcal{M}_1}(\omega_{t_1}, \omega_{t_3}) \leq \varepsilon, \quad d_{\mathcal{M}_1}(\omega_{t_1}, \omega_{t_2}) \geq 2\varepsilon \quad \text{and} \quad d_{\mathcal{M}_1}(\omega_{t_2}, \omega_{t_3}) \geq 2\varepsilon, \tag{6.5}$$

where $d_{\mathcal{M}_1}$ denotes some suitable metric inducing weak convergence on $\mathcal{M}_1(\mathbb{R}^d)$.

Definition 6.8 (Flickering random measures). Let $\{Z[k], k \in \mathbb{N}\}$ be a family of measure-valued processes on $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$. If there exists an $\varepsilon > 0$ and a sequence $\delta_k \downarrow 0$, such that

$$\liminf_{k\to\infty} \mathbb{P}\big\{Z[k] \text{ exhibits an } \varepsilon\text{-}\delta_k\text{-spark in } [0,T]\big\} > 0,$$

then we say that $\{Z[k], k \in \mathbb{N}\}$ is a family of *"flickering random measures"*.

The space-time scaling family of many generalized Fleming-Viot processes satisfies this definition, as is shown in [4]:

Lemma 6.9 (Generalized Fleming-Viot processes as flickering random measures). *If $\alpha < 2$ and $\Lambda((0,1]) > 0$, there exists $\varepsilon > 0$ such that*

$$\liminf_{k\to\infty} \mathbb{P}\big\{Y^{\Lambda,\Delta_\alpha}[k] \textit{ exhibits an } \varepsilon\text{-}(1/k)\textit{-spark in } [0,T]\big\} > 0.$$

Hence, the scaling family $\{Y^{\Lambda,\Delta_\alpha}[k], k \geq 1\}$ is a family of "flickering random measures" *with $\delta_k = 1/k, k \in N$.*

Intuitive picture: The behaviour of $\{Y^{\Lambda,\Delta_\alpha}[k]\}$, leading to an ε–$(1/k)$-spark described by condition (6.5) arises in the lookdown-framework as follows: At time t_1, suppose that $Y^{\Lambda,\Delta_\alpha}[k]$ is (almost) concentrated in a small ball with (random) center x, say. At time t_2, suddenly a fraction ε of the total mass appears in a remote ball with center y, and vanishes almost instantaneously, i.e., by time t_3. This happens precisely when, due to the stable motion, a particle with a level higher than 1 jumps to the remote position y and almost instantaneously gives birth to a positive fraction of the population (which happens if it is the lowest particle involved in a birth event). At this time, the spark appears. However, it typically fades away almost instantaneously, since the following birth events are more likely to involve lower levels, which typically are still based inside in the small ball with center x, say at time t_3.

This behaviour leads to the fact that the modulus of continuity $w'(\cdot,\delta,T)$ of the processes $Y^{\Lambda,\Delta_\alpha}[k]$, does not become small as $\delta \to 0$, contradicting relative compactness of distributions on $D_{[0,\infty)}(\mathcal{M}_1(\mathbb{R}^d))$. Intuitively, at each infinitesimal "spark", a limiting process is neither left- nor right-continuous. One should also note that such "sparks" typically appear at a dense set of time-points. For details, see [4].

The situation is different if $\Lambda = c\delta_0$ for some $c > 0$ and $\alpha < 2$. Here, each $Y^{c\delta_0,\Delta_\alpha}[k]$ a.s. has continuous paths, so that any limit in Skorohod's J_1-topology would necessarily have continuous paths. However, the f.d.d. limit $\{\delta_{B_t^{(\alpha)}}, t \geq 0\}$ has no continuous modification. Intuitively, there is no "flickering", but an "afterglow" effect: From time to time, the very fertile level-1 particle jumps some long distance, and then founds a large family, so that the population quickly becomes essentially a Dirac measure at this point, while at the same time the rest of the population (continuously) "fades away".

7. Remarks on the properties of the support of Beta-Fleming-Viot processes

Although much is known about the support of the $(2, d, \beta)$-superprocess, and in particular the classical Dawson-Watanabe process (which is an interesting fractal in dimensions $d \geq 2$, see, e.g., [30], [10], [14], [33], [7], and the references therein), this does not seem to be the case for the closely related Beta$(2-\beta, \beta)$-Fleming-Viot process with underlying Brownian motion, and certainly not for even more general Fleming-Viot processes.

One of the reasons is that Fleming-Viot processes (other than the exceptional case of the normalized Neveu superprocess, cf. Proposition 4.7) are *not* infinitely divisible and therefore do not allow a simple decomposition into independent clusters.

In [11], Dawson and Hochberg show that that for fixed times the support of the Fleming-Viot process in dimension d has Hausdorff dimension not greater than two. More than a decade later, Reimers [36] proves that the Fleming-Viot process is supported for all times on a random set of Hausdorff dimension at most two, using non-standard techniques.

To overcome the technical obstacles preventing stronger results, we suggest to use the results developed in Section 4 and Section 5, and in particular Theorem 4.1, to gain information about the support of the Beta-Fleming-Viot process from known results about the corresponding $(2, d, \beta)$-superprocesses. In particular, bounds on the behaviour of the time change (4.6) will be required. Note that also Snake techniques, developed, e.g., in [25] and [26], may be applied, where now the underlying motion needs to be assumed to be time-inhomogeneous, i.e., time-changed according to (4.6).

We conclude with a quick result on the non-compactness of the support of the Beta$(1, 1)$-Fleming-Viot process:

Proposition 7.1. *The support of the Beta*$(1, 1)$*-Fleming-Viot process* (*with underlying Brownian motion*) *propagates instantaneously throughout* $\mathbb{R}^d$*, i.e., for each time* $t > 0$

$$\operatorname{supp}\left(Y_t^{1,\Delta}\right) = \mathbb{R}^d \qquad a.s.$$

This follows immediately from Proposition 4.7 and the fact that the Neveu superprocess propagates instantaneously in $\mathbb{R}^d$ even if the underlying motion is Brownian, see [19, Proposition 14]. Further details may be found in [37].

The lookdown-construction and the duality to the Bolthausen-Sznitman coalescent provide an intuitive explanation for this. In fact, a large class of generalized Fleming-Viot processes does not have the compact support property even if the underlying motion is Brownian and the initial state has compact support. Recall Definition 3.4. The following observation can be found in [4].

Proposition 7.2. *Let* $\Lambda \in \mathcal{M}_F([0, 1])$ *and let* $\{\Pi_t^\Lambda\}$ *be the corresponding* Λ*-coalescent. If* $\{\Pi_t^\Lambda\}$ *does not come down from infinity, then, for each* $t > 0$,

$$\operatorname{supp}\left(Y_t^{\Lambda,\Delta}\right) = \mathbb{R}^d \qquad a.s.$$

This can be understood easily: If Λ-coalescent does not come down from infinity (a necessary and sufficient condition for this can be found in [39]), it either has a positive fraction of singleton classes (so-called “dust”), or countably many families with strictly positive asymptotic mass adding up to one (so-called “proper frequencies”), cf. [32], Lemma 25. Using the pathwise embedding of the standard Λ-coalescent in the Fleming-Viot process provided by the modified lookdown construction (see (5.7) above) we see that in the first case, the positive fraction of singletons contributes an α-heat flow component to $Y_t^{\Lambda,\Delta_\alpha}$, whereas in the latter case there are infinitely many independent families of strictly positive mass, so that by the Borel-Cantelli Lemma any given open ball in $\mathbb{R}^d$ will be charged almost surely.

Acknowledgment

The author wishes to thank Matthias Birkner for his fruitful and friendly collaboration over several years on several of the topics of this article and Matthias Steinrücken for helpful comments.

References

[1] Bertoin, J.; Le Gall, J.F.: Stochastic flows associated to coalescent processes, *Probab. Theory Related Fields* **126**, no. 2, 261–288, (2003).

[2] Billingsley, P.: *Convergence of probability measures*, Wiley (1968).

[3] Birkner, M.; Blath, J.: Measure-valued diffusions, general coalescents and population genetic inference, in: *Trends in Stochastic Analysis*, LMS 353, Cambridge University Press, 329–363, (2009).

[4] Birkner, M.; Blath, J.: Rescaled stable generalised Fleming-Viot processes: Flickering random measures, *WIAS Preprint* **1252**, 14 pages, submitted, (2008).

[5] Birkner, M.; Blath, J.; Capaldo, M.; Etheridge, A.; Möhle, M.; Schweinsberg, J.; Wakolbinger, A.: α-stable branching and β-coalescents, *Electronic Journal of Probability* **10,** Paper no. 9, 303–325, (2005).

[6] Birkner, M.; Blath, J.; Möhle, M.; Steinrücken, M.; Tams, J.: A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks, *ALEA Lat. Am. J. Probab. Math. Stat.* **6**, 25–61, (2009).

[7] Blath, J.; Mörters, P.: Thick points of super-Brownian motion. *Probab. Theory Related Fields* **131**, no. 4, 604–630, (2005)

[8] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Probab.* **6**, 260–290 (1974).

[9] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. *Adv. Appl. Probab.* **7**, 264–282 (1975).

[10] Dawson, D.: *Measure-valued Markov processes*, École d'été de Probabilités de Saint Flour XXI, Lecture Notes in Mathematics **1541** pp 1–260, Springer-Verlag, (1993).

[11] Dawson, D.; Hochberg, K.: Wandering random measures in the Fleming-Viot Model. *Ann. Probab.* **10**, no. 3, 554–580, (1982).

[12] Donnelly, P.; Kurtz, T.: A countable representation of the Fleming-Viot measure-valued diffusion, *Ann. Appl. Probab.* **24**, 698–742, (1996).

[13] Donnelly, P.; Kurtz, T.: Particle representations for measure-valued population models, *Ann. Probab.* **27**, no. 1, 166–205, (1999).

[14] E.A. Perkins and S.J. Taylor. The multifractal structure of super-Brownian motion. *Ann. Inst. Henri Poincaré* **34**, 97–138, (1998).

[15] Etheridge, A.M.: *An Introduction to Superprocesses.* AMS University Lecture Series, Vol. 20, (2000).

[16] Etheridge, A.; March, P.: A note on superprocesses, *Probab. Theory Rel. Fields* **89**, 141–148, (1991).

[17] Ethier, S.; Kurtz, T.: *Markov Processes: Characterization and Convergence*, Wiley, New York, (1986).

[18] Feller, W.: *Introduction to Probability Theory and its Applications,* Volume II, Wiley, (1966).

[19] Fleischmann, K.; Sturm, A.: A super-stable motion with infinite mean branching, *Ann. Inst. H. Poincaré Probab. Statist.* **40**, no. 5, 513–537, (2004).

[20] Fleischmann, K.; Wachtel, V.: Large scale localization of a spatial version of Neveu's branching process. *Stochastic Process. Appl.* **116**, no. 7, 983–1011, (2006).

[21] Fleming, W.; Viot, M.: Some measure-valued Markov processes in population genetics theory. *Indiana Univ. Math. J.* **28** 817–843, (1979).

[22] Hiraba, S.: Jump-type Fleming-Viot processes. *Adv. in Appl. Probab.* **32**, no. 1, 140–158, (2000).

[23] Kingman, J.F.C.: The coalescent. *Stoch. Proc. Appl.* **13**, 235–248, (1982).

[24] Kurtz, T.G., Rodriguez, E.: Poisson representations of branching Markov and measure-valued branching processes. *Preprint*, (2009).

[25] Le Gall, J.-F.: Brownian excursions, trees and measure-valued branching processes. *Ann. Probab.* **19**, 1399–1439, (1991).

[26] Le Gall, J.-F.; Le Jan, Y.: Branching processes in Lévy processes: the exploration process. *Ann. Probab.* **26**, no. 1, 213–252, (1998).

[27] Le Gall, J.-F.; Perkins, E.A.: The Hausdorff measure of the support of two-dimensional super-Brownian motion. *Ann. Probab.* **23**, 1719–1747, (1995).

[28] Mörters, P.: How fast are the particles of super-Brownian motion? *Probab. Theory Relat. Fields* **121**, 171–197, (2001).

[29] Möhle, M.; Sagitov, S.: A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29**, 1547–1562, (2001).

[30] Perkins, E.A: The Hausdorff measure of the closed support of super-Brownian motion. *Ann. Inst. Henri Poincaré Probab. Statist.* **25**, 205–224, (1989).

[31] Perkins, E.: *Conditional Dawson-Watanabe processes and Fleming-Viot processes*, Seminar in Stochastic Processes, Birkhäuser, pp. 142–155, (1991).

[32] Pitman, J.: Coalescents with multiple collisions, *Ann. Probab.* **27**, no. 4, 1870–1902, (1999).

[33] Perkins, E.A: Dawson-Watanabe Superprocesses and Measure-valued Diffusions. *Springer Lecture Notes in Mathematics* **1781**, (2002).

[34] Pitman, J.: Combinatorial stochastic processes. *Lecture Notes in Mathematics*, **1875**. Springer-Verlag, (2006).

[35] Prohorov, Y.V.: Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* **1**, 157–214, (1956).

[36] Reimers, M.: A new result on the support of the Fleming-Viot process, proved by nonstandard construction. *Stochastics Stochastics Rep.* **44**, no. 3-4, 213–223, (1993).

[37] Ruscher, J.: Properties of Superprocesses and Interacting particle Systems, Diplomarbeit, TU Berlin, (2009).

[38] Sagitov, S.: The general coalescent with asynchronous mergers of ancestral lines, *J. Appl. Probab.* **26**, 1116–1125, (1999).

[39] Schweinsberg, J.: A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Comm. Probab.* **5**, 1–11, (2000).

[40] Schweinsberg, J.: Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5**, Paper no. 12, (2000).

[41] Skorohod, A.V.: Limit theorems for stochastic processes. *Theory Probab. Appl.* **1**, 261–290, (1956).

[42] Tavaré, S.: Ancestral Inference in Population Genetics. *Springer Lecture Notes* **1837**, 2001.

[43] Watanabe, S.: A limit theorem of branching processes and continuous state branching processes. *J. Math. Kyoto Univ.* **8**, 141–167, (1968).

[44] Yosida, K.: *Functional Analysis*, Grundlehren der mathematischen Wissenschaften **123**, Springer-Verlag, (1965).

Jochen Blath
Technische Universität Berlin
Institut für Mathematik
D-10623 Berlin, Germany
e-mail: blath@math.tu-berlin.de

Progress in Probability, Vol. 61, 197–224

Random Maps and Their Scaling Limits

Grégory Miermont

Abstract. We review some aspects of scaling limits of random planar maps, which can be considered as a model of a continuous random surface, and have driven much interest in the recent years. As a start, we will treat in a relatively detailed fashion the well-known convergence of uniform plane trees to the Brownian Continuum Random Tree. We will put a special emphasis on the fractal properties of the random metric spaces that are involved, by giving a detailed proof of the calculation of the Hausdorff dimension of the scaling limits.

Mathematics Subject Classification (2000). 60C05, 60F17.

Keywords. Random maps, random trees, scaling limits, Brownian CRT, random metric spaces, Hausdorff dimension.

1. Introduction

A planar map is an embedding of a finite, connected graph (loops and multiple edges are allowed) into the 2-dimensional sphere. A planar map determines *faces*, which are the connected components of the complementary of the union of edges. The set of edges, vertices, and faces of the map $\mathbf{m}$ are denoted by $E(\mathbf{m}), V(\mathbf{m}), F(\mathbf{m})$. The Euler Formula asserts that

$$\#V(\mathbf{m}) - \#E(\mathbf{m}) + \#F(\mathbf{m}) = 2\,. \tag{1.1}$$

A map is said to be *rooted* if one of its oriented edges, called the root edge, is distinguished. The origin of the root edge is called the root vertex. Two (rooted) maps are systematically identified if there exists an orientation-preserving homeomorphism of the sphere that corresponds the two embedded graphs (and the roots). With these identifications, maps are combinatorial objects, which can be understood as the non-equivalent ways of gluing by pairs the edges of a finite set of polygons, so that the resulting surface is the sphere. See [50, Chapter 3] for a detailed discussion. The number of edges of a polygon is called the *degree* of the corresponding face in the map: it is the number of edges incident to the face, where edges incident to only one face are counted twice (once for each orientation).

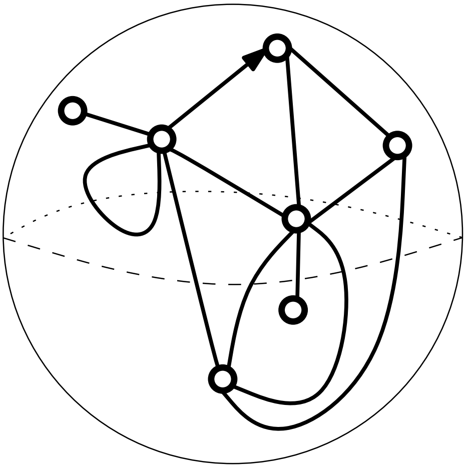

Figure 1. A rooted planar map with faces of degrees $1, 3, 4, 7$

Familiar examples of maps are triangulations, where all faces have degree 3, and quadrangulations, where faces have degree 4.

The theory of (planar) maps takes its roots in graph theory, with the 4-color theorem, and has developed considerably in other branches of mathematics. Tutte [54] founded the combinatorial study of planar maps by developing methods to solve the equations satisfied by the associated generating functions. It was then noticed by theoretical physicists, starting from 't Hooft [32] and Brézin, Parisi, Itzykson and Zuber [16], that the generating functions of maps can be interpreted as certain matrix integrals. This initiated an extremely abundant literature on (colored) graph enumeration, and has deep connections with statistical physics, representation theory and algebraic geometry, see the book by Lando and Zvonkin [35] and the recent article [30].

In the last few years, there has been also a growing interest in understanding the geometric structure of a randomly chosen map. This was partly motivated by the so-called 2-dimensional *quantum gravity theory* arising in theoretical physics, in which ill-defined measures over spaces of surfaces are considered. There are several attempts to make such theories mathematically rigorous. One of them, called Liouville theory, is to extend the language of Riemannian geometry to (very irregular) random fields, see for instance [24] for motivation. It is however not understood at present how to obtain well-defined random metric spaces with this approach.

Another approach [5, 4] is to consider discrete versions of surfaces, a role that is naturally performed by maps, and to take *scaling limits*. Scaling limits are relatively common when dealing with combinatorial aspects of probability theory: one chooses an object at random inside a class of discrete objects, and as the size of the object grows, the latter approaches, once suitably normalized, a continuous structure. The most familiar situation is to consider the Wiener measure, i.e., the law of Brownian motion, as the scaling limit of the law of the simple random walk. This legitimates viewing the Wiener measure as the "uniform measure over continuous paths". The latter has a *universal* character, in the sense that any centered random walk whose i.i.d. steps have a finite variance also admit Brownian motion as a scaling limit, according to the Donsker invariance principle. Another well-known study of scaling limits of discrete structures is that of random trees, for instance uniform random plane trees, which are known to converge to Aldous' Brownian Continuum Random Tree (CRT) [1, 2, 3]. In turn, this is a universal limit for many models of trees, e.g., arising from branching processes.

One can attempt to follow the same approach for maps: consider a large random planar map, say with uniform distribution among the set of quadrangulations of the sphere with n faces, and endow the set of its vertices with the graph distance. This means that the distance between two vertices is the minimal number of edges needed to link them. This yields a random, finite metric space. As the number of faces grows larger, the typical distances in the map are expected to grow like a power of n, which turns out to be $n^{1/4}$ as we will see below (Theorem 4.1). Therefore, one tries to understand the limiting behavior of the map where graph distances are all multiplied by $n^{-1/4}$. In some sense, it is expected that these random metric spaces converge to a limiting random surface, the so-called *Brownian map.*

This last approach turns out to be mathematically tractable thanks to powerful bijective encodings of maps, initiated by Schaeffer [52], and taking their roots in the work of Cori and Vauquelin [22] and Arquès [8]. With this method, the (hard) study of maps is amenable to the (simpler) study of certain decorated trees, in such a way that crucial geometric information of the maps, like graph distances between vertices, are encoded in a convenient way in the underlying tree structure. The extensive study of random trees in the probabilistic literature allows to understand in a detailed way the structure of large random planar maps.

This line of reasoning was first explored by Chassaing and Schaeffer [21]. This was pursued by Marckert and Mokkadem [43], who introduced a natural scaling limit for random quadrangulations, while Marckert, Miermont and Weill [42, 55, 46, 49] studied the universal aspects of these results, building on the powerful generalization of Schaeffer's bijection due to Bouttier, Di Francesco and Guitter [13]. In two important papers, Le Gall [38] and Le Gall and Paulin [39] studied in detail the fractal and topological structure of the scaling limits of random quadrangulations (and more generally 2κ-angulations). Further properties on the geodesics in the limit are known [45, 40]. Hence, at present, there is a relatively good understanding of the problem of scaling limits of planar (and non-planar) maps, even though many open questions remain, the most important (Conjecture 4.5) being a problem of identification of the limiting space. More precisely, the best that can be shown at present is, using a relative compactness argument, that scaling limits exist *up to taking extractions*, and the properties that are known up to now are valid for any scaling limit. The question of uniqueness for the scaling limit appears in Schramm [53].

In the present survey, we will introduce some of the basic and more elaborate results in this vein, focusing essentially on the particular case of random uniform quadrangulations of the sphere with n faces. It turns out to be the simplest model that can be shown to converge, up to extraction, to a limiting continuous structure. Like Brownian motion, these "Brownian map" limits have a singular, fractal structure, and we will put a special emphasis on the derivation of the Hausdorff dimension.

We also mention that a related, but different approach of random maps exists in the literature. One can also consider *local limits*, in which the convergence of arbitrary large but finite neighborhoods of the root of a large random planar map is studied. Here, the lengths are not normalized, and neighboring vertices remain at distance 1, so that the objects arising in the limit are still discrete (but infinite) structures. These infinite limiting random graphs have driven much interest [7, 6, 20, 34]. However, we will not cover this approach here.

The rest of the paper is organized as follows. Section 2 gathers the combinatorial tools that are required in the study of quadrangulations, and we give a detailed construction of the Schaeffer bijection. In Section 3, we will focus on the scaling limits of labeled trees, which are the key tools to understand scaling limits of quadrangulations. In Section 4, we discuss the construction of the scaling limits of random quadrangulations, discuss some of its most important properties and give a detailed computation of the Hausdorff dimension. Finally, Section 5 gathers developments on more general situations (more general families of maps, higher genera), some recent results, and open questions.

2. Coding quadrangulations with trees

2.1. Notations

Let

$$\mathcal{U} = \bigcup_{n\geq 0} \mathbb{N}^n ,$$

where $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}^0 = \{\varnothing\}$ by convention. We denote by $u = u_1 \ldots u_n$ a word in $\mathbb{N}^n$, and call $|u| = n$ the *height* of u. The concatenation of the words u, v is denoted by $uv = u_1 \ldots u_{|u|} v_1 \ldots v_{|v|}$. We say that u is a prefix of v, if there exists w such that $v = uw$. For $u \neq \varnothing$, the maximal strict prefix of u is called the *parent* of u. For $u, v \in \mathcal{U}$, the common prefix of u, v with maximal length is denoted by $u \wedge v$, and called the most recent common ancestor to u, v.

A *rooted plane tree* is a finite subset $\mathbf{t}$ of $\mathcal{U}$ such that

- $\varnothing \in \mathbf{t}$
- if $u \in \mathbf{t}$ and $u \neq \varnothing$, then the parent of u is in $\mathbf{t}$
- if $u \in \mathbf{t}$ and $i \geq 1$ is such that $ui \in \mathbf{t}$, then $uj \in \mathbf{t}$ for every $1 \leq j \leq i$.

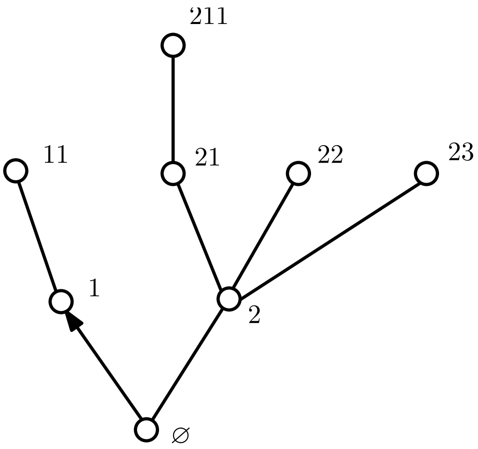

FIGURE 2. The rooted plane tree $\{\varnothing, 1, 11, 2, 21, 211, 22, 23\}$

The elements of $\mathbf{t}$ are called *vertices*. The maximal $i \geq 1$ such that $ui \in \mathbf{t}$ is called the number of children of u, and denoted by $c_u(\mathbf{t}) \geq 0$. The word $\varnothing$ is called the root vertex of $\mathbf{t}$. We let $\mathbf{T}_n$ be the set of rooted plane trees with $n+1$ vertices.

If a rooted plane tree $\mathbf{t} \in \mathbf{T}_n, n \geq 1$ is given, one can turn it into a planar map by drawing edges between each $u \in \mathbf{t}$ and its children, so that the edges between u and $u1, \ldots, uc_u(\mathbf{t})$ appear in this order when turning clockwise around u, and are followed by the edge between u and its parent whenever $u \neq \varnothing$. This embedding is naturally rooted at the oriented edge pointing from $\varnothing$ to 1, and yields a rooted planar map with one face by the Jordan Curve Theorem, since the underlying graph is connected and has no loops. These two points of view turn out to be equivalent, so that we will also refer to $\mathbf{T}_n$ as the set of trees with n edges. This is summed up in Figure 2.

A *labeled* plane tree (with n edges) is a pair of the form $(\mathbf{t}, \ell)$, where $\mathbf{t} \in \mathbf{T}_n$ and ℓ is a labeling function defined on the set of vertices of $\mathbf{t}$, with values in $\mathbb{Z}$, and such that $\ell(\varnothing) = 0$ and $|\ell(u) - \ell(ui)| \leq 1$ whenever $u \in \mathbf{t}$ and $1 \leq i \leq c_u(\mathbf{t})$. Let $\mathbb{T}_n$ be the set of labeled plane trees with n edges. The cardinality of this set equals

$$|\mathbb{T}_n| = 3^n |\mathbf{T}_n| = 3^n \frac{1}{n+1} \binom{2n}{n}. \tag{2.1}$$

Indeed, since $\ell(\varnothing) = 0$ is fixed, choosing a labeling for the tree $\mathbf{t}$ is equivalent to choosing the increments of ℓ along the n edges of $\mathbf{t}$, i.e., the quantities $\ell(ui) - \ell(u) \in \{-1, 0, 1\}$, where $u \in \mathbf{t}$ is a vertex of $\mathbf{t}$ and $1 \leq i \leq c_u(\mathbf{t})$.

2.2. The Schaeffer bijection

With every labeled tree $(\mathbf{t}, \ell)$, we want to associate a planar quadrangulation. To this end, we introduce the so-called *contour* exploration of $\mathbf{t}$. Let $\varphi(0) = \varnothing$ be the root vertex of $\mathbf{t}$, and given $\varphi(0),\ldots,\varphi(i)$ have been constructed, let $\varphi(i+1)$ be the first child of $\varphi(i)$ that does not belong to $\{\varphi(0),\ldots,\varphi(i)\}$, if any, otherwise, $\varphi(i+1)$ is the parent of $\varphi(i)$. At step $i=2n$, all vertices have been visited and $\varphi(2n)=\varnothing$. To see this, note that the oriented edges e_i pointing from $\varphi(i)$ to $\varphi(i+1)$, for $0\leq i\leq 2n-1$, are an enumeration of the $2n$ oriented edges of $\mathbf{t}$. They form a path that "wraps around" $\mathbf{t}$ in clockwise order, starting from the root. It is convenient to extend the sequence $(\varphi(i), 0\leq i\leq 2n)$ by periodicity, by letting $\varphi(i)=\varphi(i-2n)$ whenever $i>2n$.

Take a particular planar representation of the tree $\mathbf{t}$, as in Figure 3, and add an extra vertex v_*, not belonging to the union of edges of $\mathbf{t}$. This extra vertex is assigned label $\ell(v_*)=\min_{u\in\mathbf{t}}\ell(u)-1$.

Now, for every i such that $0\leq i\leq 2n-1$, we let

$$s(i)=\inf\{j\geq i:\ell(\varphi(j))=\ell(\varphi(i))-1\}\in\mathbb{Z}_+\cup\{\infty\}\,,$$

and call it the *successor* of i. Similarly, $\varphi(s(i))$ is called a successor of the vertex $\varphi(i)$. For every $0\leq i\leq 2n-1$, we then draw an *arch*, i.e., an edge between $\varphi(i)$ and the successor $\varphi(s(i))$, where by convention, $\varphi(\infty)=v_*$. Note that the number of arches drawn from a particular vertex equals the number of times when this vertex is visited in the contour exploration, which equals its degree.

We claim that it is possible to draw the arches in such a way that they do not cross other arches nor edges of $\mathbf{t}$, i.e., such that the resulting graph is a map. When we delete the interior of the edges of $\mathbf{t}$, it holds that the resulting embedded graph $\mathbf{q}$ is still a map, and in fact, a quadrangulation. We adopt a rooting convention for this map, as follows. Choose $\epsilon\in\{-1,1\}$, and consider the arch between $\varnothing=\varphi(0)$ and $\varphi(s(0))$. If $\epsilon=1$, the root is chosen to be this arch, oriented from $\varphi(0)$ to $\varphi(s(0))$, and if $\epsilon=-1$, we choose the reverse orientation. See Figure 3 for a summary of the construction.

We are now able to state the key combinatorial result. Let $\mathbf{Q}_n$ be the set of rooted planar quadrangulations with n faces. Let $\mathbf{Q}_n^*$ be the set of pairs $(\mathbf{q},v_*)$, where $\mathbf{q}\in\mathbf{Q}_n$ and $v_*\in V(\mathbf{q})$ is a distinguished vertex.

Theorem 2.1. *The previous construction yields a bijection between the set $\mathbf{Q}_n^*$, and the set $\mathbb{T}_n\times\{-1,1\}$. This construction identifies vertices of $\mathbf{t}$ with vertices of $\mathbf{q}$ different from v_*, in such a way that for every $v\in\mathbf{t}$,*

$$d_{\mathbf{q}}(v,v_*)=\ell(v)-\ell(v_*)=\ell(v)-\min_{u\in\mathbf{t}}\ell(u)-1\,,\tag{2.2}$$

where $d_{\mathbf{q}}$ is the graph distance on $V(\mathbf{q})$.

See Figure 3 for an example. The identification of vertices of $\mathbf{q}$ distinct from v_* and the vertices of the labeled tree associated with $\mathbf{q}$ is crucial, and will be systematic in the sequel. Note that the previous theorem admits as a simple corollary the computations of the cardinality of $\mathbf{Q}_n$. First note that for every rooted quadrangulation $\mathbf{q}\in\mathbf{Q}_n$, it holds that $\#V(\mathbf{q})=n+2$, by applying the Euler Formula. Now, each choice of one vertex v_* among the $n+2$ possible yields a

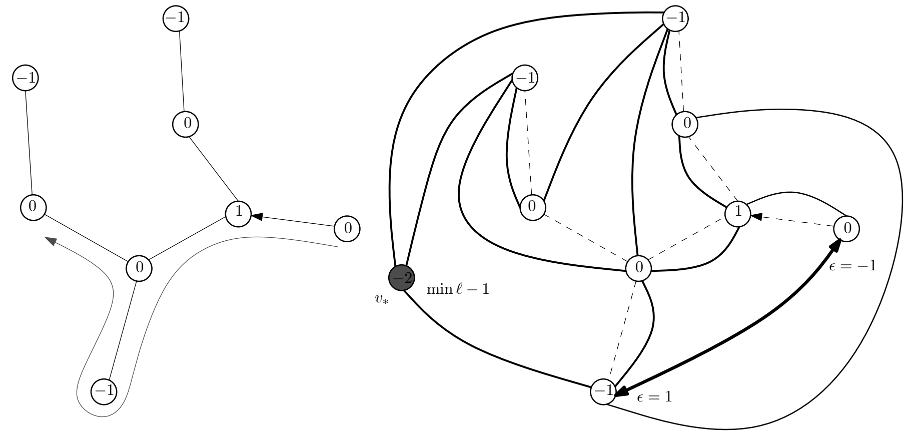

FIGURE 3. A labeled plane tree with the five first steps of the contour exploration and the associated planar quadrangulation with a distinguished vertex v_* (and with the two possible rooting choices)

different element of $\mathbf{Q}_n^*$, so combining with (2.1) yields

$$\#\mathbf{Q}_n = \frac{2}{n+2}\frac{3^n}{n+1}\binom{2n}{n}.$$

Note that the factor 2 appearing in the first term of this formula is due to the choice of ϵ in the construction. This kind of simple enumeration formula is what initially led Cori and Vauquelin [22] on the path of finding bijection between maps and labeled trees.

3. The scaling limit of labeled trees

The Schaeffer bijection leads us to consider the behavior of a uniform element of $\mathbb{T}_n$ as n gets large. We will study this problem with some detail, both because it will be crucial for the sequel, but also because it is a simple example of a derivation of a scaling limit for a random combinatorial structure, which can be seen as a sort of warm-up for the study of maps.

3.1. The Brownian CRT

To begin with, we study random trees without the labels. Let T_n be a random variable with uniform distribution in $\mathbf{T}_n$. It is a well-known fact that the typical distances between vertices of this tree are of order $n^{1/2}$. We want to rescale these distances by this factor and let n go to ∞.

3.1.1. Convergence of the contour process. Introduce the *contour process* $(C_{\mathbf{t}}(i), 0 \leq i \leq 2n)$ of $\mathbf{t} \in \mathbf{T}_n$, defined by

$$C_{\mathbf{t}}(i) = |\varphi(i)| ,$$

the height of $\varphi(i)$, where as before $\varphi(i), i \geq 0$ is the contour exploration of $\mathbf{t}$ (so that $\varphi(0) = \varphi(2n) = \varnothing$ is the root vertex). The contour process $C_{\mathbf{t}}$ is extended to a continuous function on the segment $[0, 2n]$ by linear interpolation between integer times. This yields a piecewise linear function, also known as *Harris encoding* of the tree $\mathbf{t}$. Conversely, any non-negative walk $(C(i), 0 \leq i \leq 2n)$ of duration $2n$, taking only ± 1 steps, and satisfying $C(0) = C(2n) = 0$, is the contour process of a uniquely defined element of $\mathbf{t}$.

Consequently, when T_n is a random variable uniformly distributed in $\mathbf{T}_n$, its contour process is a simple random walk with duration $2n$, conditioned to remain non-negative and to end at 0. A generalization of the Donsker invariance principle, due to Kaigh [33] shows that

$$\left(\frac{1}{\sqrt{2n}} C_{T_n}(2ns), 0 \leq s \leq 1\right) \xrightarrow[n\to\infty]{(d)} (\mathbb{e}_s, 0 \leq s \leq 1) , \tag{3.1}$$

in distribution for the uniform topology on the space $\mathcal{C}([0, 1])$ of real-valued continuous functions defined on $[0, 1]$, and where the limit is the so-called *normalized Brownian excursion*, which can be understood as an excursion away from 0 of the standard Brownian motion, conditioned to have duration 1. It can be easily defined, by scaling properties of Brownian motion, as a rescaled version of the excursion of a Brownian motion straddling 1 [51]. If $(B_s, s \geq 0)$ is a standard one-dimensional Brownian motion, and

$$g = \sup\{s \leq 1 : B_s = 0\} , \qquad d = \inf\{s \geq 1 : B_s = 0\} ,$$

then the process $\mathbb{e}$ has same distribution as

$$\frac{|B_{(d-g)s+g}|}{\sqrt{d-g}} , \quad 0 \leq s \leq 1 . \tag{3.2}$$

In terms of trees, the interpretation of (3.1) is that the process of heights in the tree, for a particle wrapping around T_n and starting from the root, converges once rescaled properly (the distances being divided by $\sqrt{2n}$) towards the contour process of a limiting structure, called the Brownian CRT [3]. The correct way to view $\mathbb{e}$ as the contour process of a tree structure is to view trees as metric spaces. Let us step back to the discrete contour process once again. Fix $0 \leq i, j \leq 2n$, let $u = \varphi(i), v = \varphi(j)$. It is then a simple exercise to check that the height of the most recent common ancestor $u \wedge v$ is equal to $\min_{i\wedge j \leq k \leq i\vee j} C_{\mathbf{t}}(k)$, moreover, any $k \in [i \wedge j, i \vee j]$ attaining this minimum is such that $\varphi(k) = u \wedge v$. As a consequence, we have a simple formula for the graph distance $d_{\mathbf{t}}$ between vertices of $\mathbf{t}$:

$$\begin{aligned} d_{\mathbf{t}}(u,v) &= |u|+|v|-2|u\wedge v| \\ &= C_{\mathbf{t}}(i)+C_{\mathbf{t}}(j)-2\min_{i\wedge j\leq k\leq i\vee j} C_{\mathbf{t}}(k) \\ &=: d^0_{\mathbf{t}}(i,j)\,. \end{aligned}$$

In the latter formula, the function $d^0_{\mathbf{t}}$ is not a distance, because it is possible to find $i \neq j$ such that $d^0_{\mathbf{t}}(i,j) = 0$. However, it is a *semi-metric*, i.e., it is non-negative, symmetric, null on the diagonal, and satisfies the triangular inequality. The quotient space obtained by identifying points at distance 0 is isometric to $(\mathbf{t}, d_{\mathbf{t}})$.

The advantage of this point of view is that it translates *verbatim* to a continuous setting. In view of (3.1) and the previous discussion, it is natural to define a "distance function" $d^0_{\mathbb{e}}$ on $[0,1]$ by letting

$$d^0_{\mathbb{e}}(s,t) = \mathbb{e}_s + \mathbb{e}_t - 2\inf_{s\wedge t\leq u\leq s\vee t} \mathbb{e}_u\,.$$

This function is the limit of the distance function $d^0_{T_n}$ in the following sense. First, we extend the distance $d^0_{\mathbf{t}}$ to $[0,2n]$ by the formula

$$\begin{aligned} d^0_{\mathbf{t}}(s,t) = {} & (\lceil s\rceil - s)(\lceil t\rceil - t)d^0_{\mathbf{t}}(\lfloor s\rfloor,\lfloor t\rfloor) + (\lceil s\rceil - s)(t-\lfloor t\rfloor)d^0_{\mathbf{t}}(\lfloor s\rfloor,\lceil t\rceil) \\ & +(s-\lfloor s\rfloor)(\lceil t\rceil - t)d^0_{\mathbf{t}}(\lceil s\rceil,\lfloor t\rfloor) + (s-\lfloor s\rfloor)(t-\lfloor t\rfloor)d^0_{\mathbf{t}}(\lceil s\rceil,\lceil t\rceil)\,, \end{aligned} \tag{3.3}$$

where by definition $\lfloor x\rfloor = \sup\{k\in\mathbb{Z}_+ : k\leq x\}$, and $\lceil x\rceil = \lfloor x\rfloor + 1$. It is easy to check that $d^0_{\mathbf{t}}$ defines a semi-metric on $[0,2n]^2$. Moreover, as a consequence of (3.1),

$$\left(\frac{d^0_{T_n}(2ns,2nt)}{\sqrt{2n}}\right)_{0\leq s,t\leq 1} \xrightarrow[n\to\infty]{(d)} (d^0_{\mathbb{e}}(s,t), 0\leq s,t\leq 1)\,, \tag{3.4}$$

in distribution for the uniform topology on $\mathcal{C}([0,1]^2)$. Again, the function $d^0_{\mathbb{e}}$ does not define a distance since, for instance, $d^0_{\mathbb{e}}(0,1) = 0$ with this definition. However, it is a semi-metric. Hence, letting $s\sim_{\mathbb{e}} t$ if $d^0_{\mathbb{e}}(s,t) = 0$, we can define a quotient metric space $T_{\mathbb{e}} = [0,1]/\sim_{\mathbb{e}}$, endowed with the quotient distance $d_{\mathbb{e}}$. This is a compact space, as it is the image of $[0,1]$ by the canonical projection, and the latter is continuous (even Hölder-continuous as we will see in Sect. 3.1.3).

Definition 3.1. The random metric space $(T_{\mathbb{e}}, \mathrm{d}_{\mathbb{e}})$ is called the Brownian Continuum Random Tree.

From the geometric point of view, this metric space indeed has a tree structure [25]. Namely, a.s. for any $a,b\in T_{\mathbb{e}}$, there is a unique injective continuous path from a to b, so that $T_{\mathbb{e}}$ has 'no loops', moreover, this path is isometric to the real segment $[0, d_{\mathbb{e}}(a,b)]$ (it is a 'geodesic'). Such spaces are usually called $\mathbb{R}$-trees [29].

3.1.2. Convergence in the Gromov-Hausdorff topology. It would be more satisfactory to view the convergence of the distance functions (3.4) as a convergence in a space of trees, rather than using the artifact of encoding spaces as quotients of $[0,1]$. This can be done by reasoning entirely in terms of metric spaces, using a topology developed around the ideas of Gromov starting in the late 1970's [31, 28].

Let $(X,d),(X',d')$ be two compact metric spaces. We let $\mathrm{d}_{\mathrm{GH}}((X,d),(X',d'))$, the Gromov-Hausdorff distance between these spaces, be the infimum of all quantities $\delta_H(\phi(X),\phi'(X'))$, taken over the set of all metric spaces (Z,δ) and isometries $\phi : X \to Z, \phi' : X' \to Z$, where δ_H denotes the Hausdorff distance between closed subsets of Z:

$$\delta_H(A,B) = \sup_{a\in A} \delta(a,B) \vee \sup_{b\in B} \delta(b,A)\,.$$

Of course, two isometric spaces will be at 'distance' 0. In fact, d_{GH} is a class function, where two spaces are identified whenever they are isometric.

Proposition 3.2. *The function* d_{GH} *is a complete, separable distance on the set of isometry classes of compact metric spaces.*

This statement is shown in [29]. We also refer to [17] for important properties of this distance. It is a simple exercise to show that the convergence (3.4) implies the following fact:

Proposition 3.3. *As $n \to \infty$, the isometry class of the random metric space $(T_n,(2n)^{-1/2}d_{T_n})$ converges in distribution to the isometry class of the Brownian CRT $(T_{\mathbb{e}},d_{\mathbb{e}})$, for the topology induced by the Gromov-Hausdorff metric on the set of isometry classes of compact metric spaces.*

3.1.3. Hausdorff dimension. As a warm-up for the later case of scaling limits of random planar maps, let us perform a Hausdorff dimension computation.

Proposition 3.4. *The Hausdorff dimension of the metric space $(T_{\mathbb{e}},d_{\mathbb{e}})$ is 2 a.s.*

This fact is well known, although complete proofs are relatively recent. Extensions to computations of exact Hausdorff measures for the Brownian CRT and other kinds of continuum trees, appear in [25, 26]. We provide a short, elementary proof, that will be useful in the analogous derivation of the dimension of scaling limits of random maps.

Proof. First of all, we use the fact that the process $\mathbb{e}$ is a.s. Hölder continuous with exponent α, for any $\alpha \in (0,1/2)$. This comes as a consequence of (3.2) and the well-known analog fact for Brownian motion. Therefore, the canonical projection $p : [0,1] \to T_{\mathbb{e}}$ is a.s. Hölder-continuous of exponent $\alpha \in (0,1/2)$. Indeed, for $s,t \in [0,1]$, choose $u \in [s\wedge t, s\vee t]$ such that $\mathbb{e}_u = \inf_{s\wedge t\leq r\leq s\vee t} \mathbb{e}_r$, and note

$$d_{\mathbb{e}}(p(s),p(t)) = \mathbb{e}_s - \mathbb{e}_u + \mathbb{e}_t - \mathbb{e}_u \leq 2\|\mathbb{e}\|_\alpha\, |s-t|^\alpha\,,$$

where

$$\|\mathbb{e}\|_\alpha := \sup_{0\leq s\neq t\leq 1} \frac{|\mathbb{e}_s - \mathbb{e}_t|}{|s-t|^\alpha}$$

is a random, a.s. finite quantity. The upper-bound $\dim_H(T_{\mathbb{e}}, d_{\mathbb{e}}) \leq \alpha^{-1}$, for any $\alpha \in (0, 1/2)$ is a direct and well-known consequence of this last fact, and letting $\alpha \to 1/2$ yields the wanted upper-bound.

Let us prove the lower bound. Let λ be the image measure of the Lebesgue measure on $[0, 1]$ by the projection p. We want to estimate the probability distribution of the distance in $T_{\mathbb{e}}$ between two λ-distributed random points. We will use the well-known fact [27, Proposition 3.4] that if U is a random variable with the uniform distribution in $[0, 1]$, independent of $\mathbb{e}$, then $2\mathbb{e}_U$ has the so-called Rayleigh distribution:

$$P(\mathbb{e}_U \geq r) = \exp(-2r^2), \qquad r \geq 0. \tag{3.5}$$

Lemma 3.5. *Let U, V be independent uniform random variables in $[0, 1]$, independent of $\mathbb{e}$. Then $d^0_{\mathbb{e}}(U, V)$ has the same distribution as $\mathbb{e}_U$.*

Proof. This lemma is a special case of a more general invariance of $T_{\mathbb{e}}$ by change of root, since $\mathbb{e}_U = d^0_{\mathbb{e}}(0, U)$ measures the distance to the special point 0, sometimes called the root of $T_{\mathbb{e}}$. The idea of its proof is simple. Let $k, l \in \{0, 1, \ldots, 2n-1\}$. The mapping from $\mathbf{T}_n$ to itself, consisting in re-rooting $\mathbf{t}$ at the edge e_k pointing from the vertex $\varphi(k)$ to $\varphi(k+1)$, is a bijection. Therefore, it leaves the uniform law on $\mathbf{T}_n$ unchanged. Now,

$$d^0_{\mathbf{t}}(k, l) = d_{\mathbf{t}}(\varphi(k), \varphi(l)) = C_{\mathbf{t}}(k) + C_{\mathbf{t}}(l) - 2 \min_{[k \wedge l, k \vee l]} C_{\mathbf{t}}.$$

In the new contour exploration of the tree $\mathbf{t}$ re-rooted at the edge e_k, the vertex $\varphi(l)$ is now visited at step $l - k$ if $k \leq l$, or $2n + l - k$ otherwise. By applying this to the uniform random variable T_n, we thus obtain that $d_{T_n}(\varphi(k), \varphi(l))$ must have the same distribution as $C_{T_n}((l-k) \vee (2n + l - k))$. Letting $k = \lfloor 2ns \rfloor, l = \lfloor 2nt \rfloor$, letting $n \to \infty$ and applying (3.1), this yields

$$d^0_{\mathbb{e}}(s, t) = \mathbb{e}_s + \mathbb{e}_t - 2 \inf_{[s \wedge t, s \vee t]} \mathbb{e} \overset{(d)}{=} \mathbb{e}((t-s) \vee (1 + t - s)),$$

for every fixed s, t. By independence, we may apply this to $s = U, t = V$, and using the fact that $(V - U) \vee (1 + V - U)$ has the same law as U, we get the result. □

We are now ready to end the proof of Proposition 3.4. First of all, with the above notations, we have, for $r \geq 0$,

$$P(d^0_{\mathbb{e}}(U, V) \leq r) = P(\mathbb{e}_U \leq r) = 1 - \exp(-2r^2) \leq 2r^2,$$

by using Lemma 3.5 and (3.5). On the other hand, the left-hand side in the above displayed expression also equals

$$E\left[\int_{T_{\mathbb{e}}} \lambda(\mathrm{d}a) \int_{T_{\mathbb{e}}} \lambda(\mathrm{d}b)\, \mathbb{1}_{\{d_{\mathbb{e}}(a,b) \leq r\}}\right] = E\left[\int_{T_{\mathbb{e}}} \lambda(\mathrm{d}a) \lambda(B_r(a))\right],$$

where $B_r(a)$ denotes the ball with radius r centered at a in $(T_{\mathbb{e}}, \mathrm{d}_{\mathbb{e}})$. This yields, for $\varepsilon > 0$,

$$E\left[\int_{T_{\mathbb{e}}} \lambda(\mathrm{d}a) \mathbb{1}_{\{\lambda(B_r(a)) \geq r^{2-\varepsilon}\}}\right] \leq 2r^2/r^{2-\varepsilon} = 2r^{\varepsilon}.$$

Applying this to $r = 2^{-k}$, we obtain that the above quantities have a finite sum as k varies in $\mathbb{N}$. The Borel-Cantelli Lemma shows that P-a.s., for λ-almost every $a \in T_{\mathbb{e}}$, there exists a (random) K such that $\lambda(B_{2^{-k}}(a)) < 2^{-(2-\varepsilon)k}$ for $k \geq K$, so that P-a.s.,

$$\limsup_{k\to\infty} \frac{\lambda(B_{2^{-k}}(a))}{2^{-(2-\varepsilon)k}} \leq 1\,, \qquad \lambda(\mathrm{d}a) - \text{a.e.}$$

We conclude that $\dim_H(T_{\mathbb{e}}, d_{\mathbb{e}}) \geq 2-\varepsilon$, by standard density theorems for Hausdorff measures [44, Theorem 6.9]. □

3.2. Scaling limit of the tree with labels

Let us now turn to the limit of a uniform random element (T_n, L_n) in the set $\mathbb{T}_n$. Such a random variable is obtained by assigning uniformly at random one of the 3^n possible label functions to a uniform random variable in $\mathbf{T}_n$, so the notation is consistent and T_n has the same distribution as in the previous section.

To understand how the labels behave, let us condition on $T_n = \mathbf{t}$ and choose $u = \varphi(i), v = \varphi(j) \in \mathbf{t}$. Let $u(0), \ldots, u(|u|)$, resp. $v(0), \ldots, v(|v|)$ be the two paths of vertices starting from the root of $\mathbf{t}$ and respectively ending at u, v, going upwards in the tree. These two sequences are equal up to the step $|u \wedge v| = \min_{[i\wedge j, i\vee j]} C_{\mathbf{t}}$ when the most recent common ancestor to u, v is reached. Now, the two sequences $(\ell(u(k)), 0 \leq k \leq |u|)$ and $(\ell(v(k)), 0 \leq k \leq |v|)$ both start from 0, share common values up to step $|u \wedge v|$, and then evolve independently from $\ell(u_{|u\wedge v|})$, moreover, their individual distributions are those of a random walk with uniform step distribution in $\{-1, 0, 1\}$, which has variance 2/3. By the central limit theorem, it is to be expected that $\ell(u(k))$ approximates a standard Gaussian distribution in the scale $\sqrt{2k/3}$. More precisely, we expect $(\ell(u(l)), 0 \leq l \leq k)$ to approximate a Brownian motion in this scale. Since by (3.1) the height $|u|$ is typically of order $\sqrt{2n}\mathbb{e}_{i/2n}$, we expect the labels to be Gaussian with variance $\mathbb{e}_{i/2n}$ in the scale $(8n/9)^{1/4}$.

It is now easy to be convinced that the following statement, taken from Chassaing and Schaeffer [21], holds. By abuse of notation we let $L_n(i) := L_n(\varphi(i))$, for every $i \geq 0$. As for contour processes, we extend L_n to a continuous function on $[0, 2n]$ by linear interpolation between integers.

Proposition 3.6. *We have the joint convergence in distribution for the uniform topology on* $\mathcal{C}([0,1])^2$*:*

$$\left(\left(\frac{1}{\sqrt{2n}} C_{T_n}(2ns) \right)_{0\leq s\leq 1}, \left(\left(\frac{9}{8n}\right)^{1/4} L_n(2ns) \right)_{0\leq s\leq 1} \right) \xrightarrow[n\to\infty]{(d)} (\mathbb{e}, Z)\,, \qquad (3.6)$$

where conditionally on $\mathbb{e}$*, the process* $(Z_s, 0 \leq s \leq 1)$ *is a centered Gaussian process with covariance* $\mathrm{Cov}\,(Z_s, Z_t) = \inf_{[s\wedge t, s\vee t]} \mathbb{e}$.

The process $(\mathbb{e}, Z)$ is sometimes referred to as the *head of the Brownian snake.* The Brownian snake [36] is a Markov process from which $(\mathbb{e}, Z)$ (which is by no means a Markov process) is obtained as a simple functional. It has an important role in the resolution of certain non-linear PDEs, a fact which we are going to need later. For now, we state two elementary, useful lemmas.

Lemma 3.7. *The process Z is a.s. Hölder continuous with any exponent $\alpha \in (0, 1/4)$.*

Lemma 3.8. *Let U be a uniform random variable in $[0, 1]$, independent of $(\mathbbm{e}, Z)$. Then $(Z_{U+s} - Z_U, 0 \leq s \leq 1)$ has the same distribution as Z, where Z_{U+s} should be understood as Z_{U+s-1} whenever $U + s > 1$.*

The proof of the first lemma is an easy application of the Kolmogorov criterion, checking that for $p > 0$,

$$E[|Z_s - Z_t|^p \,|\, \mathbbm{e}] = C_p \left(\mathbbm{e}_s + \mathbbm{e}_t - 2 \inf_{[s\wedge t, s\vee t]} \mathbbm{e}\right)^{p/2} \leq 2^{p/2} C_p \|\mathbbm{e}\|_{2\alpha} |s-t|^{\alpha p} \,,$$

where C_p is the pth moment of a standard Gaussian random variable, and using the fact that $\|\mathbbm{e}\|_{2\alpha} < \infty$ a.s. for $\alpha \in (0, 1/4)$. The second lemma is a re-rooting result whose proof is analogous to that of Lemma 3.5. Details are left to the interested reader.

4. Scaling limits of random planar quadrangulations

Let us now draw consequences of Sections 2 and 3 in the context of random maps.

4.1. Limit laws for the radius and the profile

Let $\mathbf{q} \in \mathbf{Q}_n$ be a rooted planar quadrangulation, and v be a vertex of $\mathbf{q}$. As before, let $d_{\mathbf{q}}$ denote the graph distance on the set of vertices of $\mathbf{q}$. We define the *radius* of $\mathbf{q}$ seen from v as

$$\mathcal{R}(\mathbf{q}, v) = \max_{u \in V(\mathbf{q})} d_{\mathbf{q}}(u, v) \,,$$

and the *profile* of $\mathbf{q}$ seen from v as the sequence

$$I_{\mathbf{q},v}(k) = \mathrm{Card}\,\{u \in V(\mathbf{q}) : d_{\mathbf{q}}(u, v) = k\} \,, \qquad k \geq 0$$

which measures the ‘volumes’ of the spheres centered at v in the graph metric. The latter can be seen as a measure on $\mathbb{Z}_+$ with total volume $n + 2$.

Theorem 4.1. *Let Q_n be a random variable with uniform distribution in $\mathbf{Q}_n$, and conditionally on Q_n, let v_* be uniformly chosen among the $n + 2$ vertices of Q_n. Let also $(\mathbbm{e}, Z)$ denote the head of the Brownian snake, as in the previous section.*

(i) *We have*

$$\left(\frac{9}{8n}\right)^{1/4} \mathcal{R}(Q_n, v_*) \xrightarrow[n\to\infty]{(d)} \sup Z - \inf Z \,.$$

(ii) *If v_{**} is another uniform vertex of Q_n chosen independently of v_*,*

$$\left(\frac{9}{8n}\right)^{1/4} d_{Q_n}(v_*, v_{**}) \xrightarrow[n\to\infty]{(d)} \sup Z \,.$$

(iii) *Finally, the following convergence in distribution holds for the weak topology on probability measures on* $\mathbb{R}_+$*:*

$$\frac{I_{Q_n,v_*}((8n/9)^{1/4}\cdot)}{n+2} \xrightarrow[n\to\infty]{(d)} \mathcal{I}\,,$$

where $\mathcal{I}$ *is the occupation measure of* Z *above its infimum, defined as follows: for every non-negative, measurable* $g : \mathbb{R}_+ \to \mathbb{R}_+$*,*

$$\langle \mathcal{I}, g\rangle = \int_0^1 \mathrm{d}s\, g(Z_s - \inf Z)\,.$$

The points (i) and (iii) are due to Chassaing and Schaeffer [21], and (ii) is due to Le Gall [37], although these references state these properties in a slightly different context, namely, in the case where v_* is the root vertex rather than a uniformly chosen vertex. This indicates that as $n \to \infty$, the root vertex plays no particular role.

Proof. We give the proof of (i) and (ii), as they are going to be the most useful. Let (T_n, L_n, ϵ) be the labeled tree associated with (Q_n, v_*) by Schaeffer's bijection (Theorem 2.1), so that (T_n, L_n) is uniform in $\mathbb{T}_n$. By (2.2), the radius of Q_n viewed from v_* then equals as $\max L_n - \min L_n + 1$. The result (i) follows immediately from this and Proposition 3.6. As for (ii), it is clear that we may in fact assume that v_{**} is uniform among the n vertices of Q_n that are distinct from v_* and the root vertex. These are identified with the set $T_n \setminus \{\varnothing\}$. Now, letting U be a uniform random variable in $[0, 1]$, independent of the contour process C_{T_n}, we let $\langle U\rangle_{T_n} = \lceil 2nU\rceil$ if C_{T_n} has slope $+1$ at U, and $\langle U\rangle_{T_n} = \lfloor 2nU\rfloor$ otherwise. One can check as an exercise that $\varphi(\langle U\rangle_{T_n})$ is uniform among the n vertices of T_n distinct from the root vertex, while $|U - \langle U\rangle_{T_n}/2n| \leq 1/2n$. Together with Proposition 3.6, this entails that $(9/8n)^{1/4}(L_n(v_{**}) - \min L_n + 1)$, which equals $d_{Q_n}(v_*, v_{**})$, converges in distribution to $Z_U - \inf Z$. By Lemma 3.8, this has the same distribution as $-\inf Z$, or as $\sup Z$, by an obvious symmetry property. □

4.2. Convergence as a metric space

We would like to be able to understand the full scaling limit picture for random maps, in a similar fashion as it was done for trees, where we showed, relying on the basic result (3.1), that the distances in discrete trees, once rescaled by $\sqrt{2n}$, converge to the distances in the continuum random tree $(\mathcal{T}_{\mathbb{e}}, d_{\mathbb{e}})$. We thus ask if there is an analog of the CRT, that arises as the limit of the properly rescaled metric spaces (Q_n, d_{Q_n}). In view of Theorem 4.1, the correct normalization for the distance should be $n^{1/4}$.

Assume that (T_n, L_n) is uniform in $\mathbb{T}_n$, let ϵ be uniform in $\{-1, 1\}$, independent of (T_n, L_n), and let Q_n be the random uniform quadrangulation with n faces and with a uniformly chosen vertex v_*, obtained from (T_n, L_n, ϵ) by Schaeffer's bijection. Here we follow Le Gall [38][1]. By the usual identification, the set $\{\varphi(i), i \geq 0\}$ of vertices of T_n explored in contour order, is understood as the set $V(Q_n)\setminus\{v_*\}$. Define a semi-metric on $\{0, \dots, 2n\}$ by letting $d_n(i,j) = d_{Q_n}(\varphi(i), \varphi(j))$. The quotient of this metric space obtained by identifying i, j whenever $d_n(i,j) = 0$ is isometric to $(V(Q_n) \setminus \{v_*\}, d_{Q_n})$. A major problem is that $d_n(i,j)$ is not a simple functional of (C_{T_n}, L_n). Indeed, the distances that we are able to handle in an easy way are distances to v_*, through the formula

$$d_{Q_n}(v_*, v) = L_n(i) - \min L_n + 1\,, \tag{4.1}$$

whenever v is a vertex visited at time i in the contour exploration of T_n. A key observation is the following.

Lemma 4.2. *Let*

$$d_n^0(i,j) = L_n(i) + L_n(j) - 2 \inf_{[i\wedge j, i\vee j]} L_n + 2\,.$$

Then it holds that $d_n \leq d_n^0$.

Proof. Assume $i < j$ without loss of generality. It is convenient to extend $L_n(i) = L_n(\varphi(i))$ to all $i \in \mathbb{Z}_+$, by continuing the contour exploration as in Section 2.2. In the construction of the quadrangulation Q_n from (T_n, L_n) via Schaeffer's bijection, successive arches are drawn between $\varphi(i), \varphi(s(i)), \varphi(s^2(i)), \dots$ until they end at v_*, and similarly for the arches drawn successively from the $\varphi(j)$.

Let $k \in \mathbb{Z}_+ \cup \{\infty\}$ be the first step after i, or equivalently after j, such that $L_n(k) = \min_{[i,j]} L_n - 1$. Then by construction, the vertex $\varphi(k)$ is the $L_n(i) - L_n(k)$th successor of $\varphi(i)$, and the $L_n(j) - L_n(k)$th successor of $\varphi(j)$. The arches between $\varphi(i), \varphi(j)$ and their respective successors, until they arrive at $\varphi(k)$, form a path in Q_n of length $L_n(i) + L_n(j) - 2L_n(k)$, which must be larger than the distance $d_n(i,j)$. □

We extend the functions d_n, d_n^0 to $[0, 2n]^2$ by adapting the formula (3.3). It is easy to check that d_n thus extended defines a semi-metric on $[0, 2n]$ (which is not the case for d_n^0 as it does not satisfy the triangular inequality), and that it still holds that $d_n \leq d_n^0$. We let

$$D_n(s,t) = \left(\frac{9}{8n}\right)^{1/4} d_n(2ns, 2nt)\,, \qquad 0 \leq s, t \leq 1\,,$$

so that the subspace $(\{i/2n, 0 \leq i \leq 2n\}, D_n)$, quotiented by points at zero D_n-distance, is isometric to $(V(Q_n) \setminus \{v_*\}, (9/8n)^{1/4} d_{Q_n})$. We define similarly the functions D_n^0 on $[0,1]^2$. Then, as a consequence of (3.6), it holds that

$$(D_n^0(s,t), 0 \leq s, t \leq 1) \xrightarrow[n\to\infty]{(d)} (D^0(s,t), 0 \leq s, t \leq 1)\,, \tag{4.2}$$

[1] At this point, it should be noted that [38, 39, 40] consider another version of Schaeffer's bijection, where no distinguished vertex v_* has to be considered. This results in considering pairs (T_n, L_n) in which L_n is conditioned to be positive. The scaling limits of such random variables are still tractable, and in fact, are simple functionals of $(\mathbb{e}, Z)$, as shown in [41, 37]. So there will be some differences with our exposition, but these turn out to be non-important.

for the uniform topology on $\mathcal{C}([0,1]^2)$, where by definition

$$D^0(s,t) = Z_s + Z_t - 2 \inf_{[s\wedge t, s\vee t]} Z .$$

We can now state

Proposition 4.3. *The family of laws of $(D_n(s,t), 0 \leq s,t \leq 1)$, as n varies, is relatively compact for the weak topology on the probability measures on $\mathcal{C}([0,1]^2)$.*

Proof. Let $s,t,s',t' \in [0,1]$. Then by a simple use of the triangular inequality, and Lemma 4.2,

$$|D_n(s,t) - D_n(s',t')| \leq D_n(s,s') + D_n(t,t') \leq D_n^0(s,s') + D_n^0(t,t') ,$$

which allows to estimate the modulus of continuity at a fixed $\delta > 0$:

$$\sup_{\substack{|s-s'|\leq\delta \\ |t-t'|\leq\delta}} |D_n(s,t) - D_n(s',t')| \leq 2 \sup_{|s-s'|\leq\delta} D_n^0(s,s') . \tag{4.3}$$

However, the convergence in distribution (4.2) entails that for every $\varepsilon > 0$,

$$\limsup_{n\to\infty} P\left(\sup_{|s-s'|\leq\delta} D_n^0(s,s') \geq \varepsilon \right) \leq P\left(\sup_{|s-s'|\leq\delta} D^0(s,s') \geq \varepsilon \right) ,$$

and the latter goes to 0 when $\delta \to 0$, with a fixed ε, by continuity of D^0 and the fact that $D^0(s,s) = 0$. Hence, taking $\eta > 0$ and letting $\varepsilon = \varepsilon_k = 2^{-k}$, we can choose $\delta = \delta_k$ (tacitly depending also on η) such that

$$\sup_{n\geq 1} P\left(\sup_{|s-s'|\leq\delta_k} D_n^0(s,s') \geq 2^{-k} \right) \leq \eta 2^{-k} , \qquad k \geq 1,$$

entailing

$$P\left(\bigcap_{k\geq 1} \left\{ \sup_{|s-s'|\leq\delta_k} D_n^0(s,s') \leq 2^{-k} \right\} \right) \geq 1 - \eta ,$$

for all $n \geq 1$. Together with (4.3), this shows that with probability at least $1-\eta$, the function D_n is in the set of functions $f : [0,1]^2 \to \mathbb{R}$ such that for every $k \geq 1$,

$$\sup_{\substack{|s-s'|\leq\delta_k \\ |t-t'|\leq\delta_k}} |f(s,t) - f(s',t')| \leq 2^{-k} ,$$

the latter set being compact by the Arzelà-Ascoli Theorem. The conclusion follows from Prokhorov's tightness Theorem [11]. □

At this point, we are allowed to say that the random distance functions D_n admit a limit in distribution, up to taking $n \to \infty$ along a subsequence:

$$(D_n(s,t), 0 \leq s,t \leq 1) \xrightarrow{(d)} (D(s,t), 0 \leq s,t \leq 1) \tag{4.4}$$

for the uniform topology on $\mathcal{C}([0,1]^2)$. In fact, we are going to need a little more than the convergence of D_n. From the relative compactness of its components, we see that the family of laws of $((2n)^{-1}C_{T_n}(2n\cdot), (9/8n)^{1/4}L_n(2n\cdot), D_n), n \geq 1$ is relatively compact in the set of probability measures on $\mathcal{C}([0,1])^2 \times \mathcal{C}([0,1]^2)$. Therefore, it is possible to choose an extraction $(n_k, k \geq 1)$ so that this triple converges in distribution to a limit, which we call $(\mathbb{e}, Z, D)$ with a slight abuse of notation. The joint convergence to the triple $(\mathbb{e}, Z, D)$ gives a coupling of D, D^0 such that $D \leq D^0$, since $D_n \leq D_n^0$ for every n.

Define a random equivalence relation on $[0,1]$ by letting $s \approx t$ if $D(s,t) = 0$. We let $M = [0,1]/\approx$ be the quotient space, endowed with the quotient distance, which we denote by d_M. Let also $s_* \in [0,1]$ be such that $Z_{s_*} = \inf Z$ (such a s_* turns out to be unique [41]). The $\approx$-equivalence class of s_* is denoted by ρ_*, it is intuitively the point of M that corresponds to v_*. The following statement is a relatively elementary corollary of (4.4).

Proposition 4.4. *The isometry class of* (M, d_M) *is the limit in distribution of the isometry class of* $(Q_n, (9/8n)^{1/4}d_{Q_n})$*, for the Gromov-Hausdorff topology, along the subsequence* $(n_k, k \geq 1)$*. Moreover, it holds that a.s. for every* $x \in M$ *and* $s \in x$ *a* $\approx$*-representative*

$$d_M(\rho_*, x) = D(s_*, s) = Z_s - \inf Z\,.$$

The last equation is of course the continuous analog of (2.2) and (4.1), and is proved by combining this with the convergence of L_n. It is tempting to call (M, d_M) the "Brownian map", although the choice of the subsequence poses a problem of uniqueness. As we see in the previous statement, only the distances to ρ_* are *a priori* defined as simple functionals of the process Z. Distances between other points in M seem to be harder to handle, and it is not known whether they are indeed uniquely defined. In the sequel, the words "Brownian map" will refer to any limit in distribution of the form (M, d_M), along some subsequence. Of course, it is natural make the following

Conjecture 4.5. *The isometry class of* $(Q_n, n^{-1/4}d_{Q_n})$ *converges in distribution for the Gromov-Hausdorff topology.*

Marckert and Mokkadem [43] and Le Gall [38] give a natural candidate for the limit (called the Brownian map in [43]) but at present, it has not been identified as the correct limit. The rest of the section is devoted to some properties that are nevertheless satisfied by *any* limit of the form (M, d_M) as appearing in Proposition 4.4, along some subsequence.

4.3. Hausdorff dimension of the limit space

The goal of this section is to prove the following result, due to Le Gall [38].

Theorem 4.6. *Almost-surely, the Hausdorff dimension of the space* (M, d_M) *is equal to* 4.

This fact takes its historic roots in the Physics literature [4]. We are going to present a proof that is slightly simpler than that of [38], in the sense that it does not rely on the precise estimates on the behavior of the Brownian snake near its minimum that are developed in [41]. Rather, it relies on more classical properties of the Brownian snake and its connection to PDEs. We are going to need the following formula, of a Laplace transform kind, due to Delmas [23].

Proposition 4.7. *It holds that*

$$\int_0^\infty \frac{\mathrm{d}r}{2\sqrt{2\pi r^3}}\left(1 - e^{-\lambda r}P(\sup Z \le r^{-1/4})\right) = \sqrt{\frac{\lambda}{2}}\left(3\coth^2((2\lambda)^{1/4}) - 2\right).$$

Proving this formula would fall way beyond the scope of the present paper, so we take this for granted. It is interesting to note that this formula for the law of $\sup Z$, which according to Theorem 4.1 (ii) measures the distance between two uniformly chosen points in a large random quadrangulation, is intimately connected to formulas appearing in the Physics literature back in the 1990's, see [5], or [4, Chapter 4.7], under the name of *two-point function.* These were derived using direct counting arguments on maps, without mentioning labeled trees or the random variable $\sup Z$ itself. See also [12] for derivations of this formula using the language of labeled trees, relying on discrete computations and scaling limit arguments.

Corollary 4.8. *There exists a finite constant $K > 0$ such that $P(\sup Z \le r) \le Kr^4$ for every $r \ge 0$.*

Proof. In this proof, the numbers K_1, K_2 denote positive, finite, universal constants. By differentiating twice the formula of Proposition 4.7, and by an elementary (but tedious) computation, we find, as $\lambda \to 0$,

$$\int_0^\infty \sqrt{r}e^{-\lambda r}P(\sup Z \le r^{-1/4})\mathrm{d}r = K_1\lambda^{-1/2} + o(\lambda^{-1/2}).$$

Note that the differentiation under the integral sign is licit in this situation. Changing variables $s = r^{-1/4}$ yields

$$\int_0^\infty s^{-7}e^{-\lambda/s^4}P(\sup Z \le s)\mathrm{d}s \le K_2\lambda^{-1/2}, \tag{4.5}$$

for $\lambda > 0$. Introducing the function $F(x) = \int_x^\infty u^{-7}e^{-1/u^4}\mathrm{d}u$ for $x \ge 0$, note that this function is positive, decreasing to 0 as $x \to \infty$, so that $\mathbb{1}_{[0,1]}(x) \le F(1)^{-1}F(x)$ for every $x \ge 0$. This yields, using (4.5) in the last step,

$$\begin{aligned}
P(\sup Z \le r) &\le F(1)^{-1}E[F(\sup Z/r)] \\
&= F(1)^{-1}\int_0^\infty u^{-7}e^{-1/u^4}P(\sup Z \le ru)\mathrm{d}u \\
&= F(1)^{-1}r^6\int_0^\infty s^{-7}e^{-r^4/s^4}P(\sup Z \le s)\mathrm{d}s \\
&\le F(1)^{-1}K_2r^6(r^4)^{-1/2} = Kr^4,
\end{aligned}$$

as wanted. □

Lemma 4.9. *Let U, V be independent uniform random variables on* $[0,1]$*, independent of D. Then $D(U,V)$ has the same distribution as* $\sup Z$.

Proof. This statement is of course reminiscent of (ii) in Theorem 4.1, and is in some sense a continuum analog. It is similar to Lemma 3.5 as well, and is also proved using a re-rooting argument. Let U, V be as in the statement. Define U_n as follows: with probability $n/(n+2)$, we let $U_n = \langle U\rangle_{T_n}$, as in the proof of Theorem 4.1, and with equal probability $1/(n+2)$, we let $U_n = *$ or $U_n = 0$. Define V_n similarly. By convention, let $\varphi(*) = v_*$. Then the vertex $\varphi(U_n)$ of Q_n, is uniformly chosen in Q_n. Obviously, $d_{Q_n}(\varphi(U_n), \varphi(V_n))$ has the same distribution as $d_{Q_n}(v_*, \varphi(U_n))$. The first random variable equals $D_n(\langle U\rangle_{T_n}/2n, \langle V\rangle_{T_n}/2n)$ with probability going to 1 as $n \to \infty$, and by (4.4) this converges to $D(U,V)$ in distribution. On the other hand, $d_{Q_n}(v_*, \varphi(U_n))$ converges in distribution to $\sup Z$ by (ii) in Theorem 4.1. □

We are now able to prove Theorem 4.6. The scheme of the proof is very similar to that of Proposition 3.4. First of all, the canonical projection $\pi : [0,1] \to M = [0,1]/\approx$ is a.s. Hölder-continuous of index $\alpha \in (0, 1/4)$, since

$$d_M(\pi(s), \pi(t)) \leq D^0(s,t) \leq 2\|Z\|_\alpha |s-t|^\alpha ,$$

by definition of D^0 and where $\|Z\|_\alpha < \infty$ a.s. by Lemma 3.7. This implies that (M, d_M) has Hausdorff dimension at most 4.

For the lower-bound, we introduce the image measure μ of Lebesgue measure on $[0,1]$ by π, and note that if $\mathcal{B}_r(a)$ denotes the ball of radius r centered at a in the space (M, d_M), and with the same notation as in Lemma 4.9,

$$E\left[\int_M \mu(\mathrm{d}a)\mu(\mathcal{B}_r(a))\right] = P(D(U,V) \leq r) = P(\sup Z \leq r) \leq K r^4 ,$$

using Lemma 4.9 and Corollary 4.8, for any $r \geq 0$. From there, the conclusion follows by taking the exact same steps as in the proof of the lower-bound in Proposition 3.4.

4.4. Topology of the limit space

In the previous section, we showed that, even though the scaling limit of uniform random quadrangulations is not yet proved to be uniquely defined, forcing us to consider appropriate extractions, any limit along such an extraction has Hausdorff dimension 4, a.s. Several other features of the limiting map can be studied in a similar way. In particular, Le Gall [38] identifies the topology of (M, d_M):

Theorem 4.10. *The metric d_M a.s. induces the quotient topology of* $[0,1]/\approx$.

In a subsequent work, Le Gall and Paulin identify the topology of (M, d_M) by establishing the following result [39].

Theorem 4.11. *The space (M, d_M) is a.s. homeomorphic to the 2-dimensional sphere.*

This shows that the limiting space of uniform random quadrangulation is a topological surface, as was expected by physicists. To prove this, one first uses a description of M as a quotient of the CRT $T_{\mathbb{e}}$ rather than $[0,1]$. More precisely, it is easy to see that the function D is a class function of $T_{\mathbb{e}}$, meaning that $D(s,t) = D(s',t')$ for every $s \sim_{\mathbb{e}} s', t \sim_{\mathbb{e}} t'$. Hence, one can see D as a function on $T_{\mathbb{e}}$, and take the alternative definition $M = T_{\mathbb{e}}/\approx$ (instead of $M = [0,1]/\approx$) where, with the obvious abuse of notations, we write $a \approx b$ if and only if $D(a,b) = 0$. The space M is endowed with the image of the semi-metric D under the canonical projection. Theorem 4.10 depends on a careful description of identified points of $T_{\mathbb{e}}$, in a way that mimics, in a continuous framework, the addition of arches to the tree T_n in the Schaeffer bijection. In turn, the tree $T_{\mathbb{e}}$ and the identifications induced by the relation $\approx$ are viewed as a pair of geodesic laminations of the hyperbolic disk. The proof of Theorem 4.11 then rests on a theorem by Moore on quotients of the sphere, and is developed in an entirely "continuum" framework.

In Miermont [48], an alternative proof of Theorem 4.11 is provided, relying on a strengthening of the Gromov-Hausdorff convergence [9] that allows to conserve topological properties of approximating spaces in the limit. It relies on proving the non-existence of small bottlenecks, i.e., of cycles with diameter $o(n^{1/4})$ separating Q_n into two parts that are of diameters $\Omega(n^{1/4})$.

5. Developments

5.1. Universal aspects of the scaling limit

It is a natural question to ask whether the results discussed above are robust, and in particular, to see if similar results hold when quadrangulations are replaced by more general maps. In fact, all the results of [38, 39] are stated and proved in the more general setting of uniform planar 2κ-angulations, meaning that all faces have the same degree equal to 2κ, where κ is an integer larger than or equal to 2.

It has been shown in a series of papers by Marckert, Miermont and Weill [42, 46, 55, 49] with increasing generality, that results similar to Theorem 4.1 are in fact true for much more general models of maps, namely, maps with a so-called *Boltzmann distribution*. This study is allowed by the bijective encodings of general planar maps, that were studied by Bouttier, Di Francesco and Guitter [13], which generalizes the Schaeffer bijection of Sect. 2. This is at the cost of considering labeled trees with a more complicated structure than elements of $\mathbb{T}_n$, but whose enumeration and probabilistic study is still tractable, using some technology on spatial multitype branching processes, developed in [42, 47].

Let us focus on the simplest case [42] of bipartite planar maps, i.e., maps whose faces all have even degrees. We let $\mathcal{M}$ be the set of rooted bipartite planar maps. Let $\mathbf{w} = (w_1, w_2, \ldots)$ be a non-negative sequence of weights, such that $w_i > 0$ for at least one index $i \geq 2$. Then one can define a non-negative measure

$W_{\mathbf{w}}$ on $\mathcal{M}$ by letting

$$W_{\mathbf{w}}(\mathbf{m}) = \prod_{f \in F(\mathbf{m})} w_{\deg(f)/2} \,.$$

To motivate this definition, suppose that $w_i = w\mathbb{1}_{\{i=\kappa\}}$ for some $\kappa \geq 2$ and $w > 0$. In this case, $W_{\mathbf{w}}$ charges only 2κ-angulations, and assigns same weight w^m to all 2κ-angulations with m faces. By the Euler Formula (1.1), these have also $n = (\kappa - 1)m + 2$ vertices. Therefore, the probability distribution

$$W_{\mathbf{w}}^{(n)} := W_{\mathbf{w}}(\cdot|\mathcal{M}_n) = \frac{W_{\mathbf{w}}(\cdot \cap \mathcal{M}_n)}{W_{\mathbf{w}}(\mathcal{M}_n)} \,,$$

where $\mathcal{M}_n \subset \mathcal{M}$ is constituted of those $\mathbf{m}$ that have n vertices (assuming the latter set has positive $W_{\mathbf{w}}$-mass), is the uniform distribution on planar 2κ-angulations with n vertices.

We say that $\mathbf{w}$ is admissible if $W_{\mathbf{w}}(\mathcal{M}) < \infty$, in which case we an define the *Boltzmann probability distribution* $P_{\mathbf{w}} = W_{\mathbf{w}}/W_{\mathbf{w}}(\mathcal{M})$. We have the following simple criterion for admissibility. For $x \geq 0$, let

$$f_{\mathbf{w}}(x) = \sum_{k \geq 0} \binom{2k+1}{k} w_{k+1} x^k \in [0, \infty] \,,$$

hence defining a completely positive power series. Let $R_{\mathbf{w}}$ denote its radius of convergence.

Proposition 5.1. *The sequence* $\mathbf{w}$ *is admissible if and only if the equation*

$$f_{\mathbf{w}}(x) = 1 - 1/x \,, \qquad x > 1 \tag{5.1}$$

admits a solution. We say that $\mathbf{w}$ *is* critical *if furthermore it holds that the solution* $Z_{\mathbf{w}}$ *is unique and*

$$Z_{\mathbf{w}}^2 f'_{\mathbf{w}}(Z_{\mathbf{w}}) = 1 \,,$$

i.e., if the graphs of the functions $f_{\mathbf{w}}$ *and* $x \mapsto 1 - 1/x$ *are tangent at* $Z_{\mathbf{w}}$. *Finally, we say that* $\mathbf{w}$ *is* regular critical *if* $R_{\mathbf{w}} > Z_{\mathbf{w}}$, *in which case we define*

$$C_{\mathbf{w}} = \frac{9}{8 + 4Z_{\mathbf{w}}^3 f''_{\mathbf{w}}(Z_{\mathbf{w}})} \,. \tag{5.2}$$

Now recall the definitions of the radius and the profile of a quadrangulation seen from a particular vertex as in Section 4.1, and extend them *verbatim* to any planar map. Of course, we let $d_{\mathbf{m}}$ be the graph distance associated with a map $\mathbf{m}$.

Theorem 5.2. *Assume* $\mathbf{w}$ *is a regular critical sequence. Let* M_n *have distribution* $W_{\mathbf{w}}^{(n)}$, *where it is assumed that* n *varies in the set* $\{k \geq 1 : W_{\mathbf{w}}(\mathcal{M}_n) > 0\}$. *Conditionally on* M_n, *let* v_*, v_{**} *be two vertices of* M_n *chosen uniformly at random.*

Then,

$$
\begin{aligned}
(C_{\mathbf{w}}n)^{-1/4}\mathcal{R}(M_n, v_*) &\xrightarrow[n\to\infty]{(d)} \sup Z - \inf Z\,,\\
(C_{\mathbf{w}}n)^{-1/4} d_{M_n}(v_*, v_{**}) &\xrightarrow[n\to\infty]{(d)} \sup Z\,,\\
\frac{I_{M_n,v_*}((C_{\mathbf{w}}n)^{1/4}\cdot)}{n} &\xrightarrow[n\to\infty]{(d)} \mathcal{I}\,,
\end{aligned}
$$

where the constant $C_{\mathbf{w}}$ is defined in (5.2)*.*

This result is implicit in [42], although it is not stated in the exact form above. In the previous reference, the Boltzmann measures are defined on the set $\mathcal{M}^*$ of pointed, rooted maps, i.e., of pairs $(\mathbf{m}, v_*)$ with v_* a distinguished vertex of $\mathbf{m}$. One then defines the measure

$$
W^*_{\mathbf{w}}(\mathbf{m}, v_*) = \prod_{f\in F(\mathbf{m})} w_{\deg(f)/2}\,, \qquad (\mathbf{m}, v_*) \in \mathcal{M}^*
$$

instead of using $W_{\mathbf{w}}$ and choosing a vertex v_* at random. However, it is immediate that

$$
W^*_{\mathbf{w}}(\{(\mathbf{m}, v_*) : \mathbf{m} \in \mathcal{M}_n\}) = nW_{\mathbf{w}}(\mathcal{M}_n)\,,
$$

so that a random variable with law $W^*_{\mathbf{w}}$ conditioned on $\{(\mathbf{m}, v_*) : \mathbf{m} \in \mathcal{M}_n\}$ is the same as a random variable with law $W^{(n)}_{\mathbf{w}}$, together with a distinguished vertex chosen uniformly at random among the n possible choices. Another small difference is that in [42], it is assumed that the root edge of $\mathbf{m}$ points from a vertex u to a vertex v such that $d_{\mathbf{m}}(u, v_*) = d_{\mathbf{m}}(v, v_*)-1$. However, this restriction can be lifted by considering the involution of $\mathcal{M}^*$ that consists in inverting the orientation of the root, since the latter has no fixed points.

The idea of the proof is as follows. Using the bijections of [13], and under the hypothesis that $\mathbf{w}$ is critical, one can show that the tree encoding a random map with Boltzmann distribution $P_{\mathbf{w}}$ is the genealogy of a two-type critical branching process with spatial labels. This allows to understand their limiting behavior thanks to invariance principles for spatial multitype branching processes developed in [42]. These results are generalized in Miermont [47] to allow to treat the case of maps without restriction on the degree.

The condition of being regular critical is not easily read directly on the sequence of weights $\mathbf{w}$, so it is not clear to see *a priori* which are the sequences that are covered by Theorem 5.2. So let us discuss some examples.

5.1.1. Uniform 2κ-angulations. Fix $\kappa \geq 2$. As discussed at the beginning of Section 5.1, in the case where $w_i = w\mathbb{1}_{\{i=\kappa\}}$, for any $w > 0$, the measure $W^{(n)}_{\mathbf{w}}$ is the uniform measure over uniform 2κ-angulations with n vertices. We have

$$
f_{\mathbf{w}}(x) = w\binom{2\kappa-1}{\kappa-1}x^{\kappa-1}\,,
$$

and it is easy to see that $\mathbf{w}$ is (regular) critical if and only if

$$w = \frac{(\kappa-1)^{\kappa-1}}{\kappa^\kappa \binom{2\kappa-1}{\kappa-1}},$$

in which case

$$C_{\mathbf{w}} = \frac{9}{4\kappa}.$$

In particular, when $\kappa = 2$, we recover the case of quadrangulations of Theorem 4.1.

5.1.2. A more general example. Let us assume that

- the sequence $\mathbf{w}$ decreases fast enough so that the radius of convergence of $f_{\mathbf{w}}$ is infinite. This includes in particular the case where $\mathbf{w}$ has finite support.
- $w_1 = 0$, so that $W_{\mathbf{w}}$ does not charge maps with faces of degree 2 (note that such faces are non-important from the point of view of the graph distance).
- $w_2 < 1/12$ and there exists $i > 2$ such that $w_i > 0$, so that $W_{\mathbf{w}}$ is not supported only by the set of quadrangulations, this last case having been studied before.

We can freely change $\mathbf{w}$ into the sequence $a \bullet \mathbf{w} := (a^{i-1}w_i, i \geq 1)$, for some $a > 0$, without changing the distribution $W_{\mathbf{w}}^{(n)}$. Indeed, a simple use of the Euler Formula shows that

$$W_{a\bullet\mathbf{w}}(\mathbf{m}) = a^{\sum_{f\in F(\mathbf{m})}(\deg(f)/2-1)} W_{\mathbf{w}}(\mathbf{m}) = a^{\#V(\mathbf{m})-2} W_{\mathbf{w}}(\mathbf{m}),$$

since $\sum_{f\in F(\mathbf{m})} \deg(f)/2$ is the number of edges of $\mathbf{m}$. This has the effect of changing the function $f_{\mathbf{w}}$ to

$$f_{a\bullet\mathbf{w}}(x) = \frac{f_{\mathbf{w}}(ax)}{a} = 3w_2 x + 10aw_3x^2 + 35a^2w_4x^3 + \cdots,$$

and the latter converges to $3w_2x < x/4$ as $a \downarrow 0$. Since the graphs of the functions $x \mapsto x/4$ and $x \mapsto 1 - 1/x$ are tangent, it easily follows that there exists a unique value $a_c > 0$ such that $a_c \bullet \mathbf{w}$ is critical, and it is necessarily regular critical since $R_{\mathbf{w}} = R_{a\bullet\mathbf{w}} = \infty$ for every $a > 0$. Therefore, Theorem 5.2 applies to the conditioned measure $W_{\mathbf{w}}^{(n)} = W_{a_c\bullet\mathbf{w}}^{(n)}$, with the scaling constant $C_{a_c\bullet\mathbf{w}}$.

5.2. Beyond the radius and the profile

It is of course tempting, using Theorem 5.2 as a basis, to try and generalize the convergence theorems obtained from Section 4.2 onwards. This turns out to be possible for most of them, without much more effort.

It is indeed an easy exercise to check, along the lines explained above, that Propositions 4.3 and 4.4, and Theorem 4.6 remain true in the more general setting of Theorem 5.2, i.e., that maps with distribution $W_{\mathbf{w}}^{(n)}$ with regular critical $\mathbf{w}$, and with graph distances rescaled by $n^{1/4}$, admit scaling limits for the Gromov-Hausdorff topology, the latter having Hausdorff dimension 4. It is also to be expected that the result identifying the topology (Theorem 4.10) holds in this setting as well.

On a more speculative basis, the natural conjecture is of course that the limiting space is "always the same", i.e., does not depend on $\mathbf{w}$ up to scaling constants.

5.3. Geodesics

There has been also a recent interest in the study of the geodesic paths in discrete maps and in the Brownian map, and the bijective methods are again good enough to give a lot of information on these aspects. In Bouttier and Guitter [14], the authors discuss the existence of "truly distinct" geodesics between two typical vertices in a large random quadrangulation, by extending the Schaeffer bijection to a family of quadrangulations with a distinguished geodesic path. In a different direction, Miermont [45] shows the uniqueness of the geodesic between two typical points chosen in the scaling limit of a critical Boltzmann-distributed quadrangulation, i.e., with distribution $P_{\mathbf{w}}$ where $w_i = 12^{-1}\mathbb{1}_{\{i=2\}}$ with the notations of Sect. 5.1. The latter uses a new family of bijections inspired by the Schaeffer bijection, called k-pointed bijections, in which an arbitrary number k of vertices of the quadrangulation are distinguished instead of only one (the vertex that we called v_*). This bijection allows to study certain geometric loci of multi-pointed maps, which are variants of the Voronoi tessellation with sources at the distinguished vertices.

A recent, deep result of Le Gall [40] shows that it is in fact possible to identify *all* the geodesics from the point ρ_* in the Brownian map[2], as defined around Proposition 4.4. Among other results, it shows that any point in M is linked to ρ_* by $1, 2$ or 3 distinct geodesics, and identifies the *cut-locus* of ρ_*, i.e., the set of points linked to ρ_* by more than one geodesic. More precisely, recall that the Brownian map is a quotient of the Brownian CRT, $M = \mathcal{T}_{\mathbb{e}}/\approx$ as mentioned in Sect. 4.4. Let $\pi : \mathcal{T}_{\mathbb{e}} \to \mathcal{T}_{\mathbb{e}}/\approx$ be the canonical projection. We let $\mathrm{Sk}(\mathcal{T}_{\mathbb{e}})$ be the set of points that disconnect $\mathcal{T}_{\mathbb{e}}$. Then the cut-locus of ρ_* is exactly the set $\pi(\mathrm{Sk}(\mathcal{T}_{\mathbb{e}}))$, and moreover the restriction of π to $\mathrm{Sk}(\mathcal{T}_{\mathbb{e}})$ is a homeomorphic embedding. This nicely identifies $\mathcal{T}_{\mathbb{e}}$ not only as a convenient tool to build the Brownian map, but also as a natural geometric object associated with it. It also entails a confluence property of the geodesics, namely, any two geodesic paths emanating from ρ_* will share a common initial segment. This shows that there is essentially a unique way to leave the point ρ_* along a geodesic, suggesting that the space (M, d_M) is very rough from a metric point of view, and very far from being a smooth surface.

5.4. Multi-point functions

Theorems 4.1 and 5.2 identify the so-called *two-point* function in the Brownian map, i.e., the distribution of the distance between two uniformly chosen points. It is natural to wonder about the distribution of mutual distances

$$(D(U_i, U_j), 1 \leq i, j \leq k)$$

[2] Again, there is a difference with [40] as the latter reference considers rooted, non-pointed maps, see however Remark (i) after Theorem 1.4 therein.

between k independent uniformly chosen points $U_1, \dots, U_k$ in $[0,1]$, independent of the distance function D of (4.4). It turns out that knowing these distributions (in fact, just knowing that these distributions are uniquely defined and do not depend on the choice of the subsequence of Section 4.2) would be sufficient in addressing the uniqueness problem of the Brownian map, and getting rid of subsequences in Proposition 4.4.

In a recent paper [15], Bouttier and Guitter made an important step in this direction, by identifying the $k = 3$-point function. One of the ingredients is a careful use of the 3-pointed bijection of [45]. Unfortunately, the method does not seem to generalize to more points.

5.5. Higher genera

It is also natural to consider maps on an orientable, compact surface with genus g, i.e., a cellular embedding of a graph in the torus with g handles. The (asymptotic) enumeration of such maps has been studied, starting from work of Bender and Canfield [10], along the lines of Tutte's enumeration methods. It is also covered by the matrix integral methods we alluded to in Section 1.

It turns out that the Schaeffer bijection generalizes nicely to this setting as well, replacing labeled trees with labeled maps with one face (of same genus), as shown by Chapuy, Marcus and Schaeffer [19]. This naturally paves the way to the probabilistic exploration of these classes of maps. In particular, the essential uniqueness of geodesics can be also obtained in this setting [45], while Chapuy [18] provides a very nice closed representation for the 2-point function and the profile of these more general maps.

Acknowledgment

Thanks to the referee for carefully pointing several typos and inaccuracies in the first version of this work.

References

[1] D.J. Aldous. The continuum random tree. I. *Ann. Probab.*, 19(1):1–28, 1991.

[2] D.J. Aldous. The continuum random tree. II. An overview. In *Stochastic analysis* (*Durham,* 1990), volume 167 of *London Math. Soc. Lecture Note Ser.*, pages 23–70. Cambridge Univ. Press, Cambridge, 1991.

[3] D.J. Aldous. The continuum random tree. III. *Ann. Probab.*, 21(1):248–289, 1993.

[4] J. Ambjørn, B. Durhuus, and T. Jonsson. *Quantum geometry. A statistical field theory approach.* Cambridge Monographs on Mathematical Physics. Cambridge University Press, Cambridge, 1997.

[5] J. Ambjørn and Y. Watabiki. Scaling in quantum gravity. *Nuclear Phys. B*, 445(1):129–142, 1995.

[6] O. Angel. Growth and percolation on the uniform infinite planar triangulation. *Geom. Funct. Anal.*, 13(5):935–974, 2003.

[7] O. Angel and O. Schramm. Uniform infinite planar triangulations. *Comm. Math. Phys.*, 241(2-3):191–213, 2003.

[8] D. Arquès. Les hypercartes planaires sont des arbres très bien étiquetés. *Discrete Math.*, 58(1):11–24, 1986.

[9] E.G. Begle. Regular convergence. *Duke Math. J.*, 11:441–450, 1944.

[10] E.A. Bender and E.R. Canfield. The asymptotic number of rooted maps on a surface. *J. Combin. Theory Ser. A*, 43(2):244–257, 1986.

[11] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[12] J. Bouttier, P. Di Francesco, and E. Guitter. Geodesic distance in planar graphs. *Nuclear Phys. B*, 663(3):535–567, 2003.

[13] J. Bouttier, P. Di Francesco, and E. Guitter. Planar maps as labeled mobiles. *Electron. J. Combin.*, 11:Research Paper 69, 27 pp. (electronic), 2004.

[14] J. Bouttier and E. Guitter. Statistics in geodesics in large quadrangulations. *J. Phys. A*, 41(14):145001, 30, 2008.

[15] J. Bouttier and E. Guitter. The three-point function of planar quadrangulations. *J. Stat. Mech. Theory Exp.*, (7):P07020, 39, 2008.

[16] E. Brézin, C. Itzykson, G. Parisi, and J.B. Zuber. Planar diagrams. *Comm. Math. Phys.*, 59(1):35–51, 1978.

[17] D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[18] G. Chapuy. The structure of dominant unicellular maps, and a connection between maps of positive genus and planar labelled trees. *Probab. Theory Relat. Fields*, 2008. To appear.

[19] G. Chapuy, M. Marcus, and G. Schaeffer. A bijection for rooted maps on orientable surfaces. arXiv:0712.3649.

[20] P. Chassaing and B. Durhuus. Local limit of labeled trees and expected volume growth in a random quadrangulation. *Ann. Probab.*, 34(3):879–917, 2006.

[21] P. Chassaing and G. Schaeffer. Random planar lattices and integrated superBrownian excursion. *Probab. Theory Related Fields*, 128(2):161–212, 2004.

[22] R. Cori and B. Vauquelin. Planar maps are well labeled trees. *Canad. J. Math.*, 33(5):1023–1042, 1981.

[23] J.-F. Delmas. Computation of moments for the length of the one-dimensional ISE support. *Electron. J. Probab.*, 8:no. 17, 15 pp. (electronic), 2003.

[24] B. Duplantier and S. Sheffield. Liouville Quantum Gravity and KPZ. arXiv:08081560.

[25] T. Duquesne and J.-F. Le Gall. Probabilistic and fractal aspects of Lévy trees. *Probab. Theory Related Fields*, 131(4):553–603, 2005.

[26] T. Duquesne and J.-F. Le Gall. The Hausdorff measure of stable trees. *ALEA Lat. Am. J. Probab. Math. Stat.*, 1:393–415 (electronic), 2006.

[27] R.T. Durrett and D.L. Iglehart. Functionals of Brownian meander and Brownian excursion. *Ann. Probability*, 5(1):130–135, 1977.

[28] S.N. Evans. Snakes and spiders: Brownian motion on **R**-trees. *Probab. Theory Related Fields*, 117(3):361–386, 2000.

[29] S.N. Evans, J. Pitman, and A. Winter. Rayleigh processes, real trees, and root growth with re-grafting. *Probab. Theory Related Fields*, 134(1):81–126, 2006.

[30] I.P. Goulden and D.M. Jackson. The KP hierarchy, branched covers, and triangulations. *Adv. Math.*, 219(3):932–951, 2008.

[31] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152 of *Progress in Mathematics*. Birkhäuser Boston Inc., Boston, MA, 1999.

[32] G. 't Hooft. A planar diagram theory for strong interactions. *Nucl. Phys. B*, 72:461–473, 1974.

[33] W.D. Kaigh. An invariance principle for random walk conditioned by a late return to zero. *Ann. Probability*, 4(1):115–121, 1976.

[34] M.A. Krikun. A uniformly distributed infinite planar triangulation and a related branching process. *Zap. Nauchn. Sem. St.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*), 307 (Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 10):141–174, 282–283, 2004.

[35] S.K. Lando and A.K. Zvonkin. *Graphs on surfaces and their applications*, volume 141 of *Encyclopaedia of Mathematical Sciences*. Springer-Verlag, Berlin, 2004.

[36] J.-F. Le Gall. *Spatial branching processes, random snakes and partial differential equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1999.

[37] J.-F. Le Gall. A conditional limit theorem for tree-indexed random walk. *Stochastic Process. Appl.*, 116(4):539–567, 2006.

[38] J.-F. Le Gall. The topological structure of scaling limits of large planar maps. *Invent. Math.*, 169(3):621–670, 2007.

[39] J.-F. Le Gall and F. Paulin. Scaling limits of bipartite planar maps are homeomorphic to the 2-sphere. *Geom. Funct. Anal.*, 18(3):893–918, 2008.

[40] J.-F. Le Gall. Geodesics in large planar maps and in the Brownian map. arXiv:0804.3012.

[41] J.-F. Le Gall and M. Weill. Conditioned Brownian trees. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(4):455–489, 2006.

[42] J.-F. Marckert and G. Miermont. Invariance principles for random bipartite planar maps. *Ann. Probab.*, 35(5):1642–1705, 2007.

[43] J.-F. Marckert and A. Mokkadem. Limit of normalized random quadrangulations: the Brownian map. *Ann. Probab.*, 34(6):2144–2202, 2006.

[44] P. Mattila. *Geometry of sets and measures in Euclidean spaces*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995. Fractals and rectifiability.

[45] G. Miermont. Tessellations of random maps of arbitrary genus, to appear in *Ann. Scient. Éc. Norm. Supér.*, 2009.

[46] G. Miermont. An invariance principle for random planar maps. In *Fourth Colloquium on Mathematics and Computer Sciences CMCS'*06, Discrete Math. Theor. Comput. Sci. Proc., AG, pages 39–58 (electronic). Nancy, 2006.

[47] G. Miermont. Invariance principles for spatial multitype Galton-Watson trees. *Ann. Inst. H. Poincaré Probab. Statist.*, 44(6):1128–1161, 2008.

[48] G. Miermont. On the sphericity of scaling limits of random planar quadrangulations. *Electron. Commun. Probab.*, 13:248–257, 2008.

[49] G. Miermont and M. Weill. Radius and profile of random planar maps with faces of arbitrary degrees. *Electron. J. Probab.*, 13:no. 4, 79–106, 2008.

[50] B. Mohar and C. Thomassen. *Graphs on surfaces*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2001.

[51] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, third edition, 1999.

[52] G. Schaeffer. *Conjugaison d'arbres et cartes combinatoires aléatoires*. PhD thesis, Université Bordeaux I, 1998.

[53] O. Schramm. Conformally invariant scaling limits: an overview and a collection of problems. In *International Congress of Mathematicians. Vol. I*, pages 513–543. Eur. Math. Soc., Zürich, 2007.

[54] W.T. Tutte. A census of planar maps. *Canad. J. Math.*, 15:249–271, 1963.

[55] M. Weill. Asymptotics for rooted planar maps and scaling limits of two-type spatial trees. *Electron. J. Probab.*, 12:Paper no. 31, 862–925 (electronic), 2007.

Grégory Miermont
CNRS & DMA, École Normale Supérieure
45, rue d'Ulm
F-75230 Paris Cedex 05, France
e-mail: gregory.miermont@ens.fr

Part 5

Iterated Function Schemes and Transformations of Fractals

Progress in Probability, Vol. 61, 227–250

Transformations Between Fractals

Michael F. Barnsley

Abstract. We observe that there exists a natural homeomorphism between the attractors of any two iterated function systems, with coding maps, that have equivalent address structures. Then we show that a generalized Minkowski metric may be used to establish conditions under which an affine iterated function system is hyperbolic. We use these results to construct families of fractal homeomorphisms on a triangular subset of $\mathbb{R}^2$. We also give conditions under which certain bilinear iterated function systems are hyperbolic and use them to generate families of homeomorphisms on the unit square. These families are associated with "tilings" of the unit square by fractal curves, some of whose box-counting dimensions can be given explicitly.

Mathematics Subject Classification (2000). 37E30, 28A80, 37B10.

Keywords. Iterated function systems, symbolic dynamics, dynamical systems.

1. Introduction

In this introduction we refer to various terms, some more or less commonplace to fractal geometers, such as "iterated function system" and "attractor", and others more specialized, such as "top of an attractor" and "address structures". These terms are explained in subsequent sections of the paper.

A fractal transformation is a special kind of transformation between the attractors of pairs of iterated function systems. Its graph is the top of the attractor of an iterated function system that is defined by coupling the original pair of iterated function systems. Approximations to fractal transformations can be calculated in low-dimensional cases by means of a modified chaos game algorithm. They have applications in digital imaging, see [7] for example.

This paper concerns several topics related to the construction of fractal transformations and conditions under which they are homeomorphisms. The first main topic is fractal tops, introduced in [5] and [6]. We generalize the theory to encompass what Kigami [18] and Kameyama [14] call "topological self-similar systems".

Theorem 3.1 shows that a fractal top is a certain bijection between a shift invariant subspace of code space, called the tops code space, and the attractor of an iterated function system. It is associated with a natural dynamical system, on the attractor, that can provide information about the tops code space.

We use fractal tops to define fractal transformations and to provide conditions, related to address structures, under which they are homeomorphisms; this provides fractal homeomorphisms between the attractors of suitably matched pairs of iterated function systems. See Theorem 3.2.

In order to apply fractal homeomorphisms to digital imaging, for example, we find that we need families of iterated function systems that satisfy two conditions. First, assuming that each member of the family has a well-defined coding map and attractor, we require that the address structure of each member of the family is the same, so that Theorem 3.2 can be applied. Second, we require that each member of the family indeed possesses a well-defined coding map and attractor. In particular, under what conditions does an affine iterated function system possess a unique attractor?

Consider for example the linear transformations $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $f_1(x_1, x_2) = (x_2, x_1/2)$ and $f_2(x_1, x_2) = (x_2/2, x_1)$. The eigenvalues of each transformation are all real and of magnitude less than one. Each transformation possesses a unique fixed point, the origin. But there are many different closed bounded sets $A \subset \mathbb{R}^2$ such that $A = f_1(A) \cup f_2(A)$. Consequently the affine iterated function system $(\mathbb{R}^2, f_1, f_2)$ does not possess a unique attractor. There exists no metric, compatible with the natural topology of $\mathbb{R}^2$, such that both f_1 and f_2 are contractions. In this case there does not exist a well-defined coding map.

Thus, our second main topic concerns this question: Under what conditions does there exist a metric, compatible with the natural topology of $\mathbb{R}^M$, such that a given affine iterated function system on $\mathbb{R}^M$ is contractive? We answer with the aid of the antipodal metric, introduced in Theorem 4.1. This leads us to the following construction. Let $\mathcal{K} \subset \mathbb{R}^M$ be a convex body. We will say that two distinct points l, m, both belonging to the boundary of $\mathcal{K}$, are antipodal when there are two disjoint parallel support hyperplanes of $\mathcal{K}$, one that contains l and one that contains m. We will also say that two distinct points p, q, both belonging to the boundary of $\mathcal{K}$, are diametric when their distance apart maximizes the distance between pairs of distinct points p', q' in $\mathcal{K}$ such that $q' - p'$ is parallel to $q - p$. The key observation, Theorem 4.2, is that the set of antipodal pairs of points is the same as the set of diametric pairs of points of $\mathcal{K}$. We say that a transformation $f : \mathbb{R}^M \to \mathbb{R}^M$ is non-antipodal with respect to $\mathcal{K}$ when $f(\mathcal{K}) \subset \mathcal{K}$ and f maps each antipodal pair into a pair of points that are not antipodal. Then a corollary of Theorem 4.4 implies that, for any iterated function system $(\mathbb{R}^M, f_1, f_2, \ldots, f_N)$ of affine transformations, each of which is non-antipodal with respect to $\mathcal{K}$, there exists a metric, Lipshitz equivalent to the Euclidean metric, such that all of the f_ns are contractions. Such systems possess a well-defined coding map and attractor. The converse statement is provided by Theorem 4.6 and is the subject of a separate paper, [1].

In Section 5 we present families of affine iterated function systems that both illustrate and apply the theory developed in Sections 2, 3, and 4. We use Theorem 4.4 to prove that all the functions in the families are contractive with respect to the antipodal metric and that, for each family, the address structure is constant, so that Theorem 3.2 can be applied. We describe the resulting families of homeomorphisms, from a triangular region to itself, and relate them to Kameyama metrics, [14]. In particular, Theorem 5.3 states that there exists a metric, compatible with the Euclidean metric, with respect to which certain affine IFSs are IFSs of similitudes.

In Theorem 6.1 in Section 6 we give sufficient conditions under which certain bilinear iterated function systems are hyperbolic. Then we use such IFSs to construct a family of homeomorphisms on the unit square in $\mathbb{R}^2$. This example involves a "tiling" of the unit square by 1-variable fractal interpolation functions. A closed form expression for some related box-counting dimensions is provided. In this way we obtain some information about the smoothness of fractal homeomorphisms.

2. Some kinds of iterated function systems and attractors

2.1. Iterated function system with a coding map

Let $N \geq 1$ be a fixed integer. Let $(\mathbb{X}, d_{\mathbb{X}})$ be a nonempty complete metric space. Let $\mathbb{H}$ denote the set of nonempty compact subsets of $\mathbb{X}$ and let $d_{\mathbb{H}}$ denote the Hausdorff metric; then $(\mathbb{H}, d_{\mathbb{H}})$ is a complete metric space.

Let Ω denote the set of all infinite sequences of symbols $\{\sigma_k\}_{k=1}^{\infty}$ belonging to the alphabet $\{1, \ldots, N\}$. We write $\sigma = \sigma_1\sigma_2\sigma_3 \cdots \in \Omega$ to denote an element of Ω, and we write ω_k to denote the kth component of $\omega \in \Omega$. We define a metric d_Ω on Ω by $d_\Omega(\sigma, \omega) = 0$ when $\sigma = \omega$ and $d_\Omega(\sigma, \omega) = 2^{-k}$ when k is the least index for which $\sigma_k \neq \omega_k$. Then (Ω, d_Ω) is a compact metric space that we refer to as *code space*. The natural topology on Ω, induced by the metric d_Ω, is the same as the product topology that is obtained by treating Ω as the infinite product space $\{1, \ldots, N\}^\infty$.

Let $f_n : \mathbb{X} \to \mathbb{X}$, $n = 1, 2, \ldots, N$ be mappings. We refer to $(\mathbb{X}, \{f_n\}_{n=1}^N)$ as an iterated function system. Let $f_n : \mathbb{X} \to \mathbb{X}$, $n = 1, 2, \ldots, N$ be continuous and let $\pi : \Omega \to \mathbb{X}$ be a continuous mapping such that

$$\pi(\sigma) = f_{\sigma_1}(\pi(S(\sigma))) \tag{2.1}$$

for all $\sigma \in \Omega$ where $S : \Omega \to \Omega$ is the shift operator, defined by $S(\sigma) = \omega$ where $\omega_k = \sigma_{k+1}$ for $k = 1, 2, \ldots$. Then we define

$$\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi)$$

to be an *iterated function system with coding map* π. Throughout we use the abbreviation IFS to mean an iterated function system with coding map.

If $\pi(\Omega) = \mathbb{X}$ then $\mathcal{F}$ is also called a *topological self-similar system*, as introduced by Kigami [18] and by Kameyama [15], see [14].

2.2. Point-fibred, contractive, and hyperbolic IFSs

We say that the IFS $\mathcal{F}$ is *point-fibred* when it possesses a coding map given by

$$\pi(\sigma) = \lim_{k\to\infty} f_{\sigma_1} \circ f_{\sigma_2} \circ \cdots \circ f_{\sigma_k}(x) \tag{2.2}$$

where we assume that the limit exists for all $\sigma \in \Omega$, is independent of $x \in \mathbb{X}$, depends continuously on σ, and the convergence to the limit is uniform in x, for $(\sigma, x) \in \Omega \times B$, for any fixed $B \in \mathbb{H}$. It is straightforward to prove that if $\mathcal{F}$ is point-fibred then its coding map is unique.

The notion of a point-fibred iterated function system was introduced by Kieninger [17], p. 97, Definition 4.3.6; however we work in a complete metric space whereas Kieninger frames his definition in a compact Hausdorff space.

We say that the IFS $\mathcal{F}$ is *contractive* when each f_n is a contraction, namely there is a number $l_n \in [0,1)$ such that $d_{\mathbb{X}}(f_n(x), f_n(y)) \leq l_n d_{\mathbb{X}}(x,y)$ for all $x, y \in \mathbb{X}$, for all n. Then $L = \max\{l_n\}$ is called a *contractivity factor* for $\mathcal{F}$. We say that a metric $\tilde{d_{\mathbb{X}}}$ on $\mathbb{X}$ is compatible with $d_{\mathbb{X}}$ when both metrics induce the same topology on $\mathbb{X}$. We say that a metric $\tilde{d_{\mathbb{X}}}$ on $\mathbb{X}$ is Lipshitz equivalent to $d_{\mathbb{X}}$ when there exists a constant $C \geq 1$ such that $\tilde{d_{\mathbb{X}}}(x,y)/C \leq d_{\mathbb{X}}(x,y) \leq C\tilde{d_{\mathbb{X}}}(x,y)$ for all $x, y \in \mathbb{X}$. We say that $\mathcal{F}$ is *hyperbolic* if there exists a metric, Lipshitz equivalent to $d_{\mathbb{X}}$, with respect to which $\mathcal{F}$ is contractive.

When $\mathcal{F}$ is hyperbolic, its coding map is given by equation (2.2). It is straightforward to prove that any hyperbolic iterated function system is point-fibred, see for example [2] (Theorem 3), but the converse is not true: Kameyama [14] has shown that there exists an abstract point-fibred IFS, wherein $\mathbb{X} = \Omega$, that is not hyperbolic. If the IFS $\mathcal{F}$ is such that $\pi(\Omega) = \mathbb{X}$ then it is point-fibred and its coding map is given by $\{\pi(\sigma)\} = \lim_{k\to\infty} f_{\sigma_1} \circ f_{\sigma_2} \circ \cdots \circ f_{\sigma_k}(\mathbb{X})$; this is proved in section 2.4. It follows that the restriction $\mathcal{F}|_A = (A, \{f_n\}_{n=1}^N, \pi)$ of the IFS $\mathcal{F}$ to its attractor A, see below, is point-fibred. Since it is possible to construct two distinct IFSs, each with the same set of functions $\{f_n\}_{n=1}^N$, but different coding maps, there exists an IFS that is not point-fibred. Thus, the set of IFSs strictly contains the set of point-fibred IFSs which, in turn, strictly contains the set of hyperbolic IFSs.

2.3. Attractors

Let $\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi)$ be an IFS. We define the *attractor* of $\mathcal{F}$ to be

$$A = \{\pi(\sigma) : \sigma \in \Omega\} \subset \mathbb{X}.$$

Clearly $A \in \mathbb{H}$, because Ω is compact and nonempty, and $\pi : \Omega \to \mathbb{X}$ is continuous. Kameyama [14] refers to A as a *topological self-similar set*. It follows from the commutation condition (2.1) that A obeys

$$A = f_1(A) \cup f_2(A) \cup \cdots \cup f_N(A) \text{ with } A \in \mathbb{H}. \tag{2.3}$$

When $\mathcal{F}$ is point-fibred, we have

$$A = \lim_{k\to\infty} \mathcal{F}^{\circ k}(B), \tag{2.4}$$

with respect to the Hausdorff metric, for all $B \in \mathbb{H}$, where $\mathcal{F}^{\circ 0}(B) = B$, $\mathcal{F}^{\circ 1}(B) = \mathcal{F}(B)$, $\mathcal{F}^{\circ 2}(B) = \mathcal{F} \circ \mathcal{F}(B)$, and so on. Consequently, if $\mathcal{F}$ is point-fibred then its attractor A can be characterized as the unique solution of (2.3). An elegant proof of this, in the hyperbolic case, is given by Hutchinson [13], Section 3.2. He observes that a contractive IFS $\mathcal{F}$ induces a contraction $\mathcal{F} : \mathbb{H} \to \mathbb{H}$ (we use the same symbol $\mathcal{F}$ both for the IFS and the mapping) defined by $\mathcal{F}(S) = \cup f_n(B)$ for all $B \in \mathbb{H}$. See also [12] and [26]. To prove that equation (2.4) holds when $\mathcal{F}$ is point-fibred, we show that the convergence in (2.2) is uniform in $(\sigma, x) \in (\Omega, B)$, for any fixed $B \in \mathbb{H}$. Suppose the contrary. Then for some $\varepsilon > 0$, for some $B \in \mathbb{H}$, for each k, we can find $\sigma^{(k)} \in \Omega$, and $b_k \in B$ so that $d_{\mathbb{X}}(f_{\sigma^{(k)},k}(b_k), \pi\left(\sigma^{(k)}\right)) \geq \varepsilon$ for all k, where $f_{\sigma^{(k)},k} = f_{\sigma_1^{(k)}} \circ f_{\sigma_2^{(k)}} \circ \cdots \circ f_{\sigma_k^{(k)}}$. Using compactness of both B and Ω, we can find subsequences $\left\{\sigma^{(k_l)}\right\}$ and $\{b_{k_l}\}$ that converge to $\sigma \in \Omega$ and $b \in B$ respectively. Since the convergence in equation (2.2) is uniform in $(\sigma, x) \in \Omega \times B$, it follows that $f_{\sigma^{(k_l)},k_l}(b_{k_l})$ converges to $\pi(\sigma)$. Using the continuity of π and of d in both its arguments, we obtain $d_{\mathbb{X}}(\pi(\sigma), \pi(\sigma)) \geq \varepsilon$ which is a contradiction.

Note that if A is the attractor of the IFS $\mathcal{F}$ then the following diagram commutes, for $n = 1, 2, \ldots, N$,

$$\begin{tikzcd}
\Omega \arrow[r, "s_n"] \arrow[d, "\pi"'] & \Omega \arrow[d, "\pi"] \\
A \arrow[r, "f_n"'] & A
\end{tikzcd} \tag{2.5}$$

where $s_n : \Omega \to \Omega$ is the inverse shift defined by $s_n(\sigma) = \omega$ where $\omega_1 = n$ and $\omega_{k+1} = \sigma_k$ for $k = 1, 2, \ldots$. This set of assertions is equivalent to "Equation (2.1) holds for all $\sigma \in \Omega$".

2.4. When is an IFS point-fibred?

Let $\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi)$ be an IFS such that $\pi(\Omega) = \mathbb{X}$. Then $\mathcal{F}$ is point-fibred. This fact was pointed out to me by Jun Kigami after my lecture at the conference, and is contained in a remark in [19]. To prove it here, we simply note that

$$\begin{aligned}
\lim_{k\to\infty} f_{\sigma_1} \circ f_{\sigma_2} \circ \cdots \circ f_{\sigma_k}(\mathbb{X}) &= \lim_{k\to\infty} f_{\sigma_1} \circ f_{\sigma_2} \circ \cdots \circ f_{\sigma_k}(\pi(\Omega)) \\
&= \lim_{k\to\infty} \pi(s_{\sigma_1} \circ s_{\sigma_2} \circ \cdots \circ s_{\sigma_k}(\Omega)) \text{ (by (2.5))} \\
&= \pi(\sigma) \text{ for all } \sigma \in \Omega.
\end{aligned}$$

The last equality follows from the observation that the IFS $(\Omega, s_1, s_2, \ldots, s_n)$ is point-fibred with attractor Ω and coding map $\pi_\Omega : \Omega \to \Omega$ given by $\pi_\Omega(\sigma) = \sigma$ for all $\sigma \in \Omega$.

2.5. When is a point-fibred IFS hyperbolic?

Kameyama [14] has shown that there exists an IFS, wherein $\mathbb{X} = \Omega$, for which there is no metric, compatible with the original topology, with respect to which it is contractive. Inspection of this abstract IFS shows that it is point-fibred.

3. Fractal transformations

3.1. The top of a topological self-similar system

The notion of fractal tops, for hyperbolic IFSs, was introduced in [4] and developed in [5] and [6].

Let $\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi_{\mathcal{F}})$ denote an IFS, and let $A_{\mathcal{F}}$ denote its attractor. Then we define

$$\pi_{\mathcal{F}}^{-1}(\{x\}) = \{\sigma \in \Omega : \pi_{\mathcal{F}}(\sigma) = x\}$$

to be the set of *addresses* of the point $x \in A_{\mathcal{F}}$.

The following definitions and observations, which are implied by the continuity of $\pi_{\mathcal{F}} : \Omega \to A_{\mathcal{F}}$ and the commutative diagrams (2.5), generalize corresponding statements for hyperbolic IFSs. See [5], Chapter 4, and [6] for examples and discussion, in the hyperbolic case, of tops functions, top addresses, and tops code spaces.

We order the elements of Ω according to $\sigma < \omega$ iff $\sigma_k > \omega_k$,where k is the least index for which $\sigma_k \neq \omega_k$. Let $\tau_{\mathcal{F}}(x) = \sup\{\sigma \in \Omega : \pi_{\mathcal{F}}(\sigma) = x\}$ for all $x \in A_{\mathcal{F}}$. Then $\Omega_{\mathcal{F}} := \{\tau_{\mathcal{F}}(x) : x \in A_{\mathcal{F}}\} \subset \Omega$ is called the *tops code space* and $\tau : A_{\mathcal{F}} \overset{\text{onto}}{\to} \Omega_{\mathcal{F}}$ is called the *tops function*, for the IFS $\mathcal{F}$. The value $\tau_{\mathcal{F}}(x)$ is called the *top address* of $x \in A_{\mathcal{F}}$. The tops function $\tau : A_{\mathcal{F}} \to \Omega_{\mathcal{F}}$ is well defined, injective and onto, [6]. It provides a right inverse to the coding map; that is, $\pi_{\mathcal{F}} \circ \tau_{\mathcal{F}}$ is the identity on $A_{\mathcal{F}}$. The inverse function, $\tau_{\mathcal{F}}^{-1} : \Omega_{\mathcal{F}} \to A_{\mathcal{F}}$ is injective, onto, and continuous. However, $\tau_{\mathcal{F}}$ may not be continuous, [6]. Let $\overline{\tau_{\mathcal{F}}^{-1}} : \overline{\Omega}_{\mathcal{F}} \to A_{\mathcal{F}}$ denote the restriction of $\pi_{\mathcal{F}}$ to $\overline{\Omega}_{\mathcal{F}}$ (the closure of $\Omega_{\mathcal{F}}$) or, equivalently, the continuous extension of $\tau_{\mathcal{F}}^{-1}$ to $\overline{\Omega}_{\mathcal{F}}$. Then $\overline{\tau_{\mathcal{F}}^{-1}}$ is continuous and onto. The ranges of both $\overline{\tau_{\mathcal{F}}^{-1}}$ and $\tau_{\mathcal{F}}^{-1}$ are equal to $A_{\mathcal{F}}$ because $A_{\mathcal{F}}$ is closed.

Notice that (2.5) implies

$$f_n(x) = \pi_{\mathcal{F}} \circ s_n \circ \tau_{\mathcal{F}}(x) \text{ for all } x \in A_{\mathcal{F}}. \tag{3.1}$$

3.2. Symbolic dynamics

The structure of the tops code space is related to symbolic dynamics as the following theorem shows.

Theorem 3.1. *Let $\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi_{\mathcal{F}})$ be an IFS with attractor $A_{\mathcal{F}}$, and let $\Omega_{\mathcal{F}}$ be the associated tops code space. Then* (i) *$S(\Omega_{\mathcal{F}}) \subset \Omega_{\mathcal{F}}$, and* (ii) *if f_1 is injective on $A_{\mathcal{F}}$ then $S(\Omega_{\mathcal{F}}) = \Omega_{\mathcal{F}}$.*

Proof. Suppose that $\sigma \in \Omega_{\mathcal{F}}$. (i) To see that $S(\sigma) \in \Omega_{\mathcal{F}}$, suppose that there is some $\omega > S(\sigma)$ such that $\pi_{\mathcal{F}}(\omega) = \pi_{\mathcal{F}} \circ S(\sigma)$. Then

$$\pi_{\mathcal{F}}(\sigma_1\omega) = f_{\sigma_1} \circ \pi_{\mathcal{F}}(\omega) = f_{\sigma_1} \circ \pi_{\mathcal{F}} \circ S(\sigma) = \pi_{\mathcal{F}} \circ (\sigma_1 S(\sigma)) = \pi_{\mathcal{F}}(\sigma).$$

But $\sigma_1\omega > \sigma$, so this contradicts the fact that $\sigma \in \Omega_{\mathcal{F}}$. Therefore $S(\sigma)$ is the largest address of $\pi_{\mathcal{F}}(S(\sigma))$, so $S(\sigma) \in \Omega_{\mathcal{F}}$. (ii) We show that, when f_1 is invertible on

$A_{\mathcal{F}}$, we have $1\sigma \in \Omega_{\mathcal{F}}$. Suppose that $1\sigma \notin \Omega_{\mathcal{F}}$. Then there is some $\omega > 1\sigma$ such that $\pi_{\mathcal{F}}(\omega) = \pi_{\mathcal{F}}(1\sigma)$, and $\omega = 1\widetilde{\sigma}$ where $\widetilde{\sigma} > \sigma$. Then

$$f_1 \circ \pi_{\mathcal{F}}(\widetilde{\sigma}) = \pi_{\mathcal{F}}(1\widetilde{\sigma}) = \pi_{\mathcal{F}}(\omega) = \pi_{\mathcal{F}}(1\sigma) = f_1 \circ \pi_{\mathcal{F}}(\sigma),$$

so since f_1 is injective, $\pi_{\mathcal{F}}(\widetilde{\sigma}) = \pi_{\mathcal{F}}(\sigma)$, which leads to a contradiction. Hence $1\sigma \in \Omega_{\mathcal{F}}$. □

3.3. The tops dynamical system

Theorem 3.1 tells us that we can define what we call the *tops dynamical system* $T_{\mathcal{F}} : A_{\mathcal{F}} \to A_{\mathcal{F}}$ (associated with the IFS $\mathcal{F}$) by

$$T_{\mathcal{F}} = \tau_{\mathcal{F}}^{-1} \circ S \circ \tau_{\mathcal{F}}.$$

When $\tau_{\mathcal{F}}$ is continuous, the topological entropy of $T_{\mathcal{F}} : A_{\mathcal{F}} \to A_{\mathcal{F}}$ is the same as that of the shift operator acting on the tops code space $\Omega_{\mathcal{F}}$. This follows from the invariance of topological entropy under topological conjugation, see Corollary 3.1.4 on p. 109 of [16].

We can use the orbits of a tops dynamical system to calculate the tops code space: for each $x \in A_{\mathcal{F}}$ the value of $\tau_{\mathcal{F}}(x) = \sigma_1\sigma_2\ldots$ is given by

$$\sigma_k = \min\{n \in \{1, 2, \ldots, N\} : T_{\mathcal{F}}^{\circ(k-1)}(x) \in f_n(A_{\mathcal{F}})\}.$$

This formula is useful when, as in the examples illustrated in Figure 3, the sets $f_n(A_{\mathcal{F}})$ have straight edges and a simple formula for $T_{\mathcal{F}}(x)$ is available.

3.4. Transformations between attractors

Let $\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi_{\mathcal{F}})$ be an IFS with attractor $A_{\mathcal{F}} = \pi_{\mathcal{F}}(\mathbb{X})$. Similarly let $\mathcal{G} = (\mathbb{Y}, \{g_n\}_{n=1}^N, \pi_{\mathcal{G}})$ be an IFS with attractor $A_{\mathcal{G}}$. Then the corresponding *fractal transformation* $T_{\mathcal{FG}} : A_{\mathcal{F}} \to A_{\mathcal{G}}$ is defined to be

$$T_{\mathcal{FG}} = \pi_{\mathcal{G}} \circ \tau_{\mathcal{F}}.$$

The transformation $T_{\mathcal{FG}}$ depends upon the ordering of the functions in $\mathcal{F}$ and $\mathcal{G}$. For example, if $\mathcal{F} = (\mathbb{X}, f_1, f_2)$, $\mathcal{G} = (\mathbb{X}, f_2, f_1)$, then in general $T_{\mathcal{FG}}$ is not the identity map on $A_{\mathcal{F}}$.

The transformation $T_{\mathcal{FG}}$ may be characterized with the aid of the IFS with coding map

$$\mathcal{H} = (A_{\mathcal{F}} \times A_{\mathcal{G}} \times \Omega, \{k_n\}_{n=1}^N, \pi_{\mathcal{H}})$$

where $k_n(x, y, \sigma) = (f_n(x), g_n(y), s_n(\sigma))$ and the coding map is defined by $\pi_{\mathcal{H}}(\sigma) = (\pi_{\mathcal{F}}(\sigma), \pi_{\mathcal{G}}(\sigma), \sigma)$. The graph of $T_{\mathcal{FG}}$ is the same as

$$\{(x, y) : (x, y, \sigma) \in A_{\mathcal{H}}, \sigma \geq \omega \text{ for all } (x, \tilde{y}, \omega) \in A_{\mathcal{H}}\}.$$

This characterization may be used to facilitate the computation of values of $T_{\mathcal{FG}}$ when $A_{\mathcal{F}}$ and $A_{\mathcal{G}}$ are subsets of $\mathbb{R}^2$, both $\mathcal{F}$ and $\mathcal{G}$ are hyperbolic, and a chaos game type algorithm is used, see [6].

3.5. Continuity

Let $\mathcal{F} = (\mathbb{X}, \{f_n\}_{n=1}^N, \pi_{\mathcal{F}})$ denote an IFS with attractor $A_{\mathcal{F}} = \pi_{\mathcal{F}}(\mathbb{X})$. The *address structure* of the $\mathcal{F}$ is defined to be

$$\mathcal{C}_{\mathcal{F}} = \{\pi_{\mathcal{F}}^{-1}(\{x\}) \cap \overline{\Omega}_{\mathcal{F}} : x \in A_{\mathcal{F}}\} \subset 2^{\Omega}.$$

It is readily seen that $\mathcal{C}_{\mathcal{F}}$ is a partition of the closure of the tops code space. Let $\mathcal{G} = (\mathbb{Y}, \{g_n\}_{n=1}^N, \pi_{\mathcal{F}})$ denote an IFS with attractor $A_{\mathcal{G}} = \pi_{\mathcal{G}}(\mathbb{Y})$. Let $\mathcal{C}_{\mathcal{G}}$ denote the address structure of $\mathcal{G}$. We write $\mathcal{C}_{\mathcal{F}} \prec \mathcal{C}_{\mathcal{G}}$ to mean that for each $U \in \mathcal{C}_{\mathcal{F}}$ there is $V \in \mathcal{C}_{\mathcal{G}}$ such that $U \subset V$. Note that if $\mathcal{C}_{\mathcal{F}} = \mathcal{C}_{\mathcal{G}}$ then $\Omega_{\mathcal{F}} = \Omega_{\mathcal{G}}$.

Theorem 3.2. *If $\mathcal{C}_{\mathcal{F}} \prec \mathcal{C}_{\mathcal{G}}$ then the fractal transformation $T_{\mathcal{FG}} : A_{\mathcal{F}} \to A_{\mathcal{G}}$ is continuous. If $\mathcal{C}_{\mathcal{F}} = \mathcal{C}_{\mathcal{G}}$ then $T_{\mathcal{FG}}(A_{\mathcal{F}}) = A_{\mathcal{G}}$ and $T_{\mathcal{FG}}$ is a homeomorphism.*

Proof. This is essentially the same as the proof in the case of hyperbolic IFSs, see [6], because the latter relies only on the definition of $\tau_{\mathcal{F}}$ and the continuity of $\pi_{\mathcal{G}}$. □

We note that if the address structures of $\mathcal{F}$ and $\mathcal{G}$ are the same then $T_{\mathcal{FG}}$ is a homeomorphism and the dynamical systems $T_{\mathcal{F}} : A_{\mathcal{F}} \to A_{\mathcal{F}}$ and $T_{\mathcal{G}} : A_{\mathcal{G}} \to A_{\mathcal{G}}$ are topologically conjugate, with

$$T_{\mathcal{G}} = T_{\mathcal{FG}} \circ T_{\mathcal{F}} \circ T_{\mathcal{FG}}^{-1}.$$

Remark 3.3. Suppose that $\mathcal{C}_{\mathcal{F}} \prec \mathcal{C}_{\mathcal{G}}$ and that $\eta : \overline{\Omega}_{\mathcal{F}} \to \overline{\Omega}_{\mathcal{G}}$ is continuous. Suppose too that η respects the relationship $\mathcal{C}_{\mathcal{F}} \prec \mathcal{C}_{\mathcal{G}}$; that is, whenever $U \in \mathcal{C}_{\mathcal{F}}$ there is $V \in \mathcal{C}_{\mathcal{G}}$ such that $\eta(U) \subset V$. Then $F_\eta := \pi_{\mathcal{G}} \circ \eta \circ \tau_{\mathcal{F}} : A_{\mathcal{F}} \to A_{\mathcal{G}}$ is continuous. This can be proved by means of a straightforward modification to the proof of Theorem 3.2. It enables us to compute additional continuous transformations between attractors, without a lot of extra work, in applications in $\mathbb{R}^2$ where we compute fractal homeomorphisms between attractors.

Remark 3.4. If the shift map $S : \overline{\Omega}_{\mathcal{F}} \to \overline{\Omega}_{\mathcal{F}}$ respects the relationship $\mathcal{C}_{\mathcal{F}} \prec \mathcal{C}_{\mathcal{F}}$ then the tops dynamical system $T_{\mathcal{F}} : A_{\mathcal{F}} \to A_{\mathcal{F}}$ is continuous. This follows from the choices $\eta = S$ and $\mathcal{F} = \mathcal{G}$ in Remark 3.3. Examples are mentioned in Section 5, see Figure 3.

4. An affine iterated function system on a convex body is point-fibred when it is non-antipodal

Let $\mathcal{F}$ be an iterated function system of affine maps acting on $\mathbb{R}^M$. Under what conditions is $\mathcal{F}$ hyperbolic? We show that there exists a metric, compatible with the Euclidean metric, such that $\mathcal{F}$ is contractive when there exists a compact convex nonempty set $\mathcal{K}$ such that all of the maps of $\mathcal{F}$ are non-antipodal with respect to $\mathcal{K}$.

We treat $\mathbb{R}^M$ as a vector space, an affine space, and a metric space. We identify a point $x = (x_1, x_2, \ldots, x_M) \in \mathbb{R}^M$ with the vector whose coordinates are $x_1, x_2, \ldots, x_M$. We write $0 \in \mathbb{R}^M$ for the point in $\mathbb{R}^M$ whose coordinates are all zero. We write xy to denote the closed line segment with endpoints x and y. The inner product of $x, y \in \mathbb{R}^M$ is denoted by $\langle x, y\rangle$. The 2-norm of a point $x \in \mathbb{R}^M$ is $\|x\| = \sqrt{\langle x, x\rangle}$. We define $S^{M-1} = \{u \in \mathbb{R}^M : ||u|| = 1\}$. The Euclidean metric $d_E : \mathbb{R}^M \times \mathbb{R}^M \to [0, \infty)$ is defined by

$$d_E(x, y) = \|x - y\| \text{ for all } x, y \in \mathbb{R}^M.$$

Let $\mathcal{K}$ be a convex body (that is, a compact convex subset of $\mathbb{R}^M$ with nonempty interior) and let $\partial\mathcal{K}$ be the boundary of $\mathcal{K}$. Let $u \in S^{M-1}$. We define $\mathcal{L}_u = \mathcal{L}_u(\mathcal{K})$ to be the unique support hyperplane of $\mathcal{K}$ with outer normal in the direction of u. See [20], p. 14. Then $\{\mathcal{L}_u, \mathcal{L}_{-u}\}$ denotes the unique pair of distinct hyperplanes, perpendicular to u, that intersect $\partial\mathcal{K}$ but contain no points of the interior of $\mathcal{K}$. [11] refers to $\{\mathcal{L}_u, \mathcal{L}_{-u}\}$ as "the two supporting hyperplanes of $\mathcal{K}$ orthogonal to u." For $u \in \mathbb{R}^M \backslash \{0\}$ we define

$$\mathcal{A}_u = \{(l, m) \in (\mathcal{L}_u \cap \partial\mathcal{K}) \times (\mathcal{L}_{-u} \cap \partial\mathcal{K})\} \text{ and } \mathcal{A} = \cup \mathcal{A}_u.$$

We say that $(l, m) \in \mathcal{A}_u$ is an *antipodal pair* of points corresponding to the direction of u, and that $\mathcal{A} = \mathcal{A}(\mathcal{K})$ is *the set of antipodal pairs* of points of $\mathcal{K}$.

4.1. The antipodal metric

We define *the width of* $\mathcal{K}$ *in the direction of* u to be

$$\mathfrak{w}(u) = \inf\{\|l - m\| : l \in \mathcal{L}_u(\mathcal{K}), m \in \mathcal{L}_{-u}(\mathcal{K})\}, \text{ for all } u \in S^{M-1},$$

and

$$|\mathcal{K}| = \sup\{\mathfrak{w}(u) : u \in S^{M-1}\};$$

see for example [20], p. 15, and [11]. Note that $\mathfrak{w} : S^{M-1} \to \mathbb{R}$ is continuous, see [27], p. 368. Since S^{M-1} is compact it follows that $|\mathcal{K}| = \mathfrak{w}(u^*)$ for some $u^* \in S^{M-1}$.

The following metric was discovered by Ross Atkins, a student at the Australian National University. It is related to a Minkowski metric, see Corollary 4.3 below, [9], p. 21, Ex. 5, and [10], p. 100; for instance, the two metrics are the same when $\mathcal{K}$ is symmetric about 0, that is, when $x \in \mathcal{K} \Leftrightarrow -x \in \mathcal{K}$. We define the *antipodal metric* $d_{\mathcal{K}} : \mathbb{R}^M \times \mathbb{R}^M \to [0, \infty)$ by

$$d_{\mathcal{K}}(x, y) = \sup\left\{\frac{\langle (y - x), u\rangle}{\mathfrak{w}(u)} : u \in S^{M-1}\right\}$$

The maximum here is achieved at some $u^* \in S^{M-1}$ because $\langle (y - x), u\rangle / \mathfrak{w}(u)$ is a continuous mapping from S^{M-1} into $\mathbb{R}$.

Let $r > 0$ be such that there is a ball of radius r and center at $x \in \mathcal{K}$, that is contained in $\mathcal{K}$.

Theorem 4.1.

(i) *$d_{\mathcal{K}}$ is a metric on $\mathbb{R}^M$.*

(ii) *The metrics $d_{\mathcal{K}}$ and d_E on $\mathbb{R}^M$ are Lipshitz equivalent, with*

$$d_E(x, y)/|\mathcal{K}| \leq d_{\mathcal{K}}(x, y) \leq d_E(x, y)/r \textit{ for all } x, y \in \mathbb{R}^M.$$

(iii) *For all $x, y \in \mathcal{K}$*

$$d_{\mathcal{K}}(x, y) \leq 1 \textit{ with equality iff } (x, y) \in \mathcal{A}(\mathcal{K}).$$

Proof. First we prove that $d_{\mathcal{K}}$ is a metric on $\mathbb{R}^M$. (a) $d_{\mathcal{K}}$ is clearly symmetric. (b) If $x = y$ then $d_{\mathcal{K}}(x, y) = 0$. If $x \neq y$ then

$$\begin{aligned} d_{\mathcal{K}}(x,y) &= \sup\left\{\frac{\langle (y-x), u\rangle}{\mathfrak{w}(u)} \;:\; u \in S^{M-1}\right\} \\ &\geq \frac{\langle (y-x),(y-x)\rangle}{\mathfrak{w}(y-x)\,\|y-x\|} = \frac{\|y-x\|}{\mathfrak{w}(y-x)} > 0. \end{aligned}$$

We have shown that $d_{\mathcal{K}}(x, y) \geq 0$, with equality if and only if $x = y$. (c) For all $x, y, z \in \mathbb{R}^M$ we have

$$\begin{aligned} d_{\mathcal{K}}(x,y) &= \sup\left\{\frac{\langle (y-z)+(z-x), u\rangle}{\mathfrak{w}(u)} \;:\; u \in S^{M-1}\right\} \\ &\leq \sup\left\{\frac{\langle (y-z), u\rangle}{\mathfrak{w}(u)} \;:\; u \in S^{M-1}\right\} + \sup\left\{\frac{\langle (z-x), u\rangle}{\mathfrak{w}(u)} \;:\; u \in S^{M-1}\right\} \\ &= d_{\mathcal{K}}(x,z) + d_{\mathcal{K}}(z,y). \end{aligned}$$

This establishes the triangle inequality and completes the proof that $d_{\mathcal{K}}$ is a metric on $\mathbb{R}^M$. To prove (ii) we simply note that

$$\frac{\|x-y\|}{|\mathcal{K}|} \leq d_{\mathcal{K}}(x,y) \leq \frac{\|x-y\|}{r}.$$

To prove (iii) we suppose first that $(x, y) \in \mathcal{A}$. Then $xy \subset \operatorname{conv}(\mathcal{L}_v \cup \mathcal{L}_{-v})$, the convex hull of $\mathcal{L}_v \cup \mathcal{L}_{-v}$, for all $v \in S^{M-1}$. It follows that $\langle (y-x), v\rangle \leq \mathfrak{w}(v)$ for all $v \in S^{M-1}$. Hence $\langle (y-x), v\rangle / \mathfrak{w}(v) \leq 1$ for all $v \in S^{M-1}$. Also $(x, y) \in \mathcal{A}$ implies there is $u \in S^{M-1}$ such that $(x, y) \in (\mathcal{L}_u \cap \mathcal{K}) \times (\mathcal{L}_{-u} \cap \mathcal{K})$. It follows that $\langle (y-x), u\rangle = \mathfrak{w}(u)$. So

$$d_{\mathcal{K}}(x,y) = \sup\left\{\frac{\langle (y-x), v\rangle}{\mathfrak{w}(v)} \;:\; v \in S^{M-1}\right\} = \frac{\langle (y-x), u\rangle}{\mathfrak{w}(u)} = 1.$$

Now suppose $x, y \in \mathcal{K}$ but $(x, y) \notin \mathcal{A}$. Then, for each $v \in S^{M-1}$, $xy \subset \operatorname{conv}(\mathcal{L}_v \cup \mathcal{L}_{-v})$, but xy does not intersect both $\mathcal{L}_v$ and $\mathcal{L}_{-v}$.

It follows that $\langle (y-x), v\rangle / \mathfrak{w}(v) < 1$ for all $v \in S^{M-1}$. Since $d_{\mathcal{K}}(x, y) = \langle (y-x), v\rangle / \mathfrak{w}(v)$ for some $v \in S^{M-1}$ we must have $d_{\mathcal{K}}(x, y) < 1$. □

4.2. The set $\mathcal{A}$ equals the set $\mathcal{D}$, the diametric pairs of points of $\mathcal{K}$

Let $u \in S^{M-1}$. We define the *diameter of $\mathcal{K}$ in the direction of u* to be

$$\mathfrak{d}(u) = \sup\{\|x-y\| : x, y \in \mathcal{K}, x - y = \alpha u, \alpha \in \mathbb{R}\}.$$

The maximum is achieved at some pair of points belonging to $\partial\mathcal{K}$ because $\mathcal{K} \times \mathcal{K}$ is convex and compact. For $u \in \mathbb{R}^M \backslash \{0\}$ we define

$$\mathcal{D}_u = \{(p,q) \in \partial\mathcal{K} \times \partial\mathcal{K} : \mathfrak{d}(u) = \|q-p\|\} \text{ and } \mathcal{D} = \cup\mathcal{D}_u.$$

We say that $(p, q) \in \mathcal{D}_u$ is a *diametric pair* of points in the direction of u, and that $\mathcal{D}$ is *the set of diametric pairs* of points of $\mathcal{K}$.

Theorem 4.2 is probably present in the convex geometry literature, but it is not well known. For example, it is not mentioned in [20] or [23]. See also [24]. It is crucial to this work because it provides the heart of Theorem 4.4.

Theorem 4.2. *The set of antipodal pairs of points of $\mathcal{K}$ is the same as the set of diametric pairs of points of $\mathcal{K}$. That is,*

$$\mathcal{A} = \mathcal{D}.$$

Proof. See [1]. The tools used are (a) that a convex body is the intersection of all strictly convex bodies that contain it and (b) that, when $\mathcal{K}$ is strictly convex, the function $f : S^{M-1} \to S^{M-1}$ defined by $f(u) = (x_u - x_{-u})/\left\|x_u - x_{-u}\right\|$ is continuous and has the property that $< f(u), u >> 0$ for all $u \in S^{M-1}$, where $x_u \in \mathcal{L}_u \cap \partial\mathcal{K}$, and $x_{-u} \in \mathcal{L}_{-u} \cap \partial\mathcal{K}$. Hence f does not map u to $-u$ for any $x \in S^{M-1}$, from which it follows by an elementary exercise in topology (see, for example, [21], problem 10, page 367) that f has degree 1 and, in particular, is surjective. □

Let d be a metric with the properties $d(x+z, y+z) = d(x,y)$ and $d(x, (1-\lambda)x + \lambda y) = \lambda d(x,y)$ for all $x, y, z \in \mathbb{R}^M$ and all $\lambda \in [0,1]$. Then there exists a convex body $\mathcal{C}$, symmetric about the origin, such that $d = d_{\mathcal{C}}$. The set $\mathcal{C}$ is given by

$$\mathcal{C} = \{x \in \mathbb{R}^M : d(x,0) \leq 1\}.$$

See [22] pp. 31–32. In this case d is called a *Minkowski metric.*

Corollary 4.3. *Let $x, y \in \mathcal{K}$ with $x \neq y$. Then there exists $(l,m) \in \mathcal{A}(\mathcal{K})$ such that lm and xy are parallel and*

$$d_{\mathcal{K}}(x,y) = \frac{\|y-x\|}{\|m-l\|} = \frac{d_E(x,y)}{\mathfrak{d}(y-x)}.$$

In particular, $d_{\mathcal{K}}$ is a Minkowski metric; it is associated with the symmetric convex body $\mathcal{C} = \{x \in \mathbb{R}^M : d_{\mathcal{K}}(x,0) \leq 1\}$.

Proof. We can find $(l,m) \in \mathcal{D}$ such that lm and xy are parallel. By Theorem 4.2 $\mathcal{A} = \mathcal{D}$, so there is a nonzero vector v with $(l,m) \in \mathcal{A}_v$. Now, by definition,

$$d_{\mathcal{K}}(x,y) = \sup\left\{\frac{\langle (y-x), u\rangle}{\mathfrak{w}(u)} : u \in S^{M-1}\right\}.$$

We claim that the maximum occurs when $u = v$, because if $u = v$ then

$$\frac{\langle (y-x), v\rangle}{\mathfrak{w}(v)} = \frac{\langle (y-x), v\rangle}{< (m-l), v >} = \frac{\|y-x\|}{\|m-l\|},$$

and if $u \neq v$ then $\langle (y-x), u\rangle / \mathfrak{w}(u) \leq \|y-x\| / \|m-l\|$. Hence, $d_{\mathcal{K}}(x,y) = \|y-x\| / \|m-l\| = \|y-x\| / \mathfrak{d}(m-l) = d_E(x,y)/\mathfrak{d}(y-x)$. □

For example, if $\mathcal{K}$ is triangle then $d_{\mathcal{K}} = d_{\mathcal{C}}$ where $\mathcal{C}$ is hexagon, symmetric about the origin.

4.3. IFSs of non-antipodal affine transformations

We say that $f : \mathbb{R}^M \to \mathbb{R}^M$ is *non-antipodal* with respect to $\mathcal{K}$ if $f(\mathcal{K}) \subset \mathcal{K}$ and $(x, y) \in \mathcal{A}(\mathcal{K})$ implies $(f(x), f(y)) \notin \mathcal{A}(\mathcal{K})$.

Theorem 4.4. *Let $f : \mathbb{R}^M \to \mathbb{R}^M$ be affine and non-antipodal with respect to a convex body $\mathcal{K}$. Then f is a contraction with respect to $d_{\mathcal{K}}$.*

Proof. Let $(x, y) \in \mathcal{K} \times \mathcal{K}$ be given with $x \neq y$. Then by Corollary 4.3 we can find $(l, m) \in \mathcal{A}(\mathcal{K})$ such that lm is parallel to xy, and

$$d_{\mathcal{K}}(x, y) = \frac{\|y - x\|}{\|m - l\|}.$$

Now consider $(f(x), f(y)) \in \mathcal{K} \times \mathcal{K}$. If $f(x) = f(y)$ then $d(f(x), f(y) = 0 < d(x, y)$. If $f(x) \neq f(y)$ then $f(l) \neq f(m)$ and the line segments $f(x)f(y)$ and $f(l)f(m)$ are parallel, because f is affine. Also$(f(l), f(m)) \in \mathcal{K} \times \mathcal{K}$ is not an antipodal pair so, by Theorem 4.1 (iii),

$$d_{\mathcal{K}}(f(l), f(m)) < 1.$$

In fact, let $\left(\tilde{L}, \tilde{M}\right) \in \mathcal{A}(\mathcal{K})$ be such that $\tilde{L}\tilde{M}$ is parallel to both $f(l)f(m)$ and $f(x)f(y)$, again using Corollary 4.3; then we must have

$$d_{\mathcal{K}}(f(l), f(m)) = \frac{\|f(m) - f(l)\|}{\left\|\tilde{M} - \tilde{L}\right\|} < 1.$$

It follows that

$$\begin{aligned} d_{\mathcal{K}}(f(y), f(x)) &= \frac{\|f(y) - f(x)\|}{\left\|\tilde{M} - \tilde{L}\right\|} \\ &= \frac{\|f(y) - f(x)\|}{\|f(m) - f(l)\|} \frac{\|f(m) - f(l)\|}{\left\|\tilde{M} - \tilde{L}\right\|} \\ &< \frac{\|f(y) - f(x)\|}{\|f(m) - f(l)\|} = \frac{\|y - x\|}{\|m - l\|} \\ &= d_{\mathcal{K}}(y, x). \end{aligned}$$

In the penultimate line we have used the facts that f is affine and xy is parallel to lm. Hence

$$g(u) := \frac{d_{\mathcal{K}}(f(y), f(x))}{d_{\mathcal{K}}(y, x)} < 1 \text{ for all } u = \frac{y - x}{\|y - x\|} \in S^{M-1}.$$

But S^{M-1} is compact, and $g(u)$ is continuous, so there exists $v \in S^{M-1}$ for which $g(u) \leq L := g(v) < 1$, for all $u \in S^{M-1}$. □

We say that the affine iterated function system $\mathcal{F} = \left(\mathbb{R}^M, f_1, f_2, \ldots, f_N\right)$ is non-antipodal with respect to a convex body $\mathcal{K}$ when f_n is non-antipodal with respect to $\mathcal{K}$ for $n = 1, 2, \ldots, N$.

Corollary 4.5. *Let $\mathcal{F} = \left(\mathbb{R}^M, f_1, f_2, \ldots, f_N\right)$ be an iterated function system of affine transformations. If there exists a convex body $\mathcal{K} \subset \mathbb{R}^M$, with respect to which $\mathcal{F}$ is non-antipodal, then $\mathcal{F}$ is hyperbolic.*

How sharp is this result? First, observe that both of the transformations of the IFS $\{\mathbb{R}^2 : f_1(x_1, x_2) = (x_2, x_1/2),\ f_2(x_1, x_2) = (x_2/2, x_1)\}$, mentioned in the introduction, are antipodal with respect to the triangle with vertices at $(0,0)$, $(1,0)$, $(0,1)$. Also, this IFS is not point-fibred because it has more than one nonempty compact invariant set: two compact nonempty invariant sets are $\{(0,0)\}$, and $\{(0,0)\} \cup \{(1/2^n, 0) : n = 0, 1, 2, \ldots\} \cup \{(0, 1/2^n) : n = 0, 1, 2, \ldots\}$. Second, observe that the IFS $\{\mathbb{R}^2 : f_1(x_1, x_2) = (x_2, x_1/2)\}$ is non-antipodal with respect to the triangle $\mathcal{T}$ with vertices at $(0,0)$, $(1.5, 0)$, $(0,1)$, and so it is contractive with respect to the metric $d_{\mathcal{T}}$. This leads to the question: Given a point-fibred affine IFS $\mathcal{F}$, does there exist a convex closed bounded set $\mathcal{K}$ with nonempty interior, such that all of the maps of the IFS are non-antipodal with respect to $\mathcal{K}$? The answer is provided by the following theorem, which was developed with collaborators after this paper was substantially completed.

Theorem 4.6. [1] *If $\mathcal{F} = (\mathbb{R}^M, f_1, f_2, \ldots, f_N)$ is an affine iterated function system then the following statements are equivalent.*

(1) *The system $\mathcal{F}$ is hyperbolic.*
(2) *The system $\mathcal{F}$ is point-fibred.*
(3) *There exists a convex body $\mathcal{K} \subset \mathbb{R}^M$ with respect to which $\mathcal{F}$ is non-antipodal.*

Note that (1) is a metric statement, (2) is a topological statement, and (3) is a geometrical statement. See [1] for proofs of this and more, including an answer to a fundamental question of Kameyama concerning topological self-similar systems.

5. Affine IFSs associated with a triangle

We use the terminology *affine IFS* to mean an IFS of affine maps. We illustrate the theory of the preceding sections with an application in $\mathbb{R}^2$.

5.1. Non-antipodal subtriangles

We choose $\mathcal{K} = \mathcal{T}$, a filled triangle in $\mathbb{R}^2$, with strictly positive area and vertices at the points A, B, and C. Let $\mathcal{T}'$denote a triangle with vertices at the points P, Q, and R. Suppose that $\mathcal{T}' \subset \mathcal{T}$. Suppose also that the statement "both $\mathcal{T}' \cap \{P\} \neq \emptyset$ and $\mathcal{T}' \cap QR \neq \emptyset$" is not true, for all cyclic permutations PQR of ABC. Then we say that $\mathcal{T}'$ is a *non-antipodal subtriangle* of $\mathcal{T}$.

Corollary 5.1. *Let the affine IFS $\mathcal{F} = \{\mathcal{T}; f_1, f_2, \ldots, f_N\}$ be such that $f_n(\mathcal{T})$ is a non-antipodal subtriangle of $\mathcal{T}$ for each n. Then $\mathcal{F}$ is contractive with respect to $d_{\mathcal{T}}$.*

This result is useful for applications because it provides a convenient geometrical condition under which an affine IFS is point-fibred. We use it next to yield families of affine IFSs, such that each family has a constant address structure.

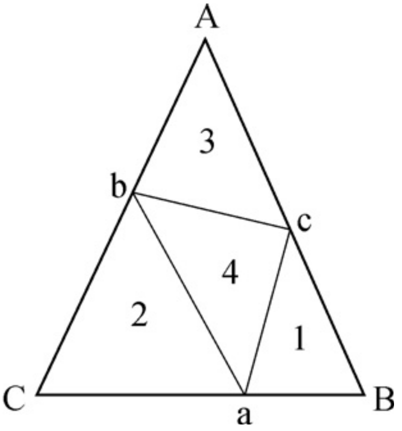

FIGURE 1. The triangles used to define the affine transformations of the IFS $\mathcal{F}_\alpha = \{\mathbb{R}^2; f_1, f_2, f_3, f_4\}$.

5.2. Families of homeomorphisms

Let $\mathcal{T}$ be a triangle with vertices A, B, C as above. Let c denote a point on the line segment AB, let a denote a point on the line segment BC, and let b denote a point on the line segment CA, such that $\{a, b, c\} \cap \{A, B, C\} = \emptyset$. Then each of the triangles caB, Cab, cAb, and cab is a non-antipodal subtriangle of $\mathcal{T}$, see Figure 1.

Let $f_1 : \mathbb{R}^2 \to \mathbb{R}^2$ denote the unique affine transformation such that

$$f_1(ABC) = caB,$$

by which we mean that f_1 maps A to c, B to a, and C to B. Using the same notation, let affine transformations f_2, f_3, and f_4 be the ones uniquely defined by

$$f_2(ABC) = Cab,\ f_3(ABC) = cAb, \quad \text{and} \quad f_4(ABC) = cab.$$

Let us write $\mathcal{F}_\alpha = \{\mathbb{R}^2; f_1, f_2, f_3, f_4\}$, where

$$\alpha = (|Bc|/|AB|, |Ca|/|BC|, |Ab|/|CA|).$$

Then, for all $\alpha \in (0,1)^3$, $\mathcal{F}_\alpha$ is contractive with respect to the metric $d_\mathcal{T}$, has constant attractor $\mathcal{T}$, and has constant address structure $\mathcal{C} := \mathcal{C}_{\mathcal{F}_\alpha}$. The latter assertion is proved in [6], Example 1. Consequently Theorem 3.2 provides a fractal homeomorphism $T_{\alpha\beta} : \mathcal{T} \to \mathcal{T}$ defined by $T_{\alpha\beta} = \pi_\beta \circ \tau_\alpha$ where $\pi_\beta := \pi_{\mathcal{F}_\beta}$ and $\pi_\alpha := \pi_{\mathcal{F}_\alpha}$, for all $\alpha, \beta \in (0,1)^3$. Note that $T_{\beta\gamma} \circ T_{\alpha\beta} = T_{\alpha\gamma}$ for all $\alpha, \beta, \gamma \in (0,1)^3$ and that $T_{\alpha\beta}$ preserves area when $\alpha = (c, c, c)$ and $\beta = (1-c, 1-c, 1-c)$ for any $c \in (0,1)$.

Since $\mathcal{F}_\alpha$ is contractive with respect to the antipodal metric $d_\mathcal{K}$, we can find a contractivity factor $L_\alpha \in (0,1)$ such that $d_\mathcal{T}(f_n^\alpha(x), f_n^\alpha(y)) \leq L_\alpha d_\mathcal{T}(x, y)$ for all n, x, y. Clearly, L_α is not a constant function of α. However, we will use the next lemma to prove that there is a metric, compatible with the Euclidean metric, with respect to which all of the f_n^αs are similitudes.

Lemma 5.2.

$$T_{\alpha\beta}(f_n^\alpha(x)) = f_n^\beta(T_{\alpha\beta}(x)) \textit{ for all } x \in \mathcal{T}, \textit{ for all } \alpha, \beta, \ n.$$

Proof. Equation (3.1) implies $f_n^\alpha(x) = \pi_\alpha \circ s_n \circ \tau_\alpha(x)$ for all $x \in \mathcal{T}$. Hence, since the tops code space $\Omega_{\mathcal{F}_\alpha}$ is independent of α, we have

$$\begin{aligned} T_{\alpha\beta}(f_n^\alpha(x)) &= T_{\alpha\beta} \circ \pi_\alpha \circ s_n \circ \tau_\alpha(x) = \pi_\beta \circ \tau_\alpha \circ \pi_\alpha \circ s_n \circ \tau_\alpha(x) \\ &= \pi_\beta \circ s_n \circ \tau_\alpha(x) = \pi_\beta \circ s_n \circ \tau_\beta \circ \pi_\beta \circ \tau_\alpha(x) \\ &= \pi_\beta \circ s_n \circ \tau_\beta \circ T_{\alpha\beta}(x) = f_n^\beta(T_{\alpha\beta}(x)), \end{aligned}$$

for all $x \in \mathcal{T}$, for all α, β, n. □

We define a family of metrics $d_{\alpha\beta}$ on $\mathcal{T}$ by

$$d_{\alpha\beta}(x, y) = d_E(T_{\alpha\beta}(x), T_{\alpha\beta}(y)).$$

For each $\alpha, \beta \in (0,1)^3$, this is indeed a metric, compatible with the Euclidean metric, because $T_{\alpha\beta} : \mathcal{T} \to \mathcal{T}$ is a homeomorphism.

Theorem 5.3. *The maps f_n^α of the IFS $\mathcal{F}_\alpha$, restricted to $\mathcal{T}$, are similitudes with scaling factor* 0.5 *with respect to the metric $d_{\alpha\hat{\beta}}$, where $\hat{\beta}$:=(0.5, 0.5, 0.5).*

Proof. Using Lemma 5.2 and the definition $d_{\alpha\hat{\beta}}$ we have

$$\begin{aligned} d_{\alpha\hat{\beta}}(f_n^\alpha(x), f_n^\alpha(y)) &= d_E(T_{\alpha\hat{\beta}}(f_n^\alpha(x)), T_{\alpha\hat{\beta}}(f_n^\alpha(y))) \\ &= d_E(f_n^{\hat{\beta}}(T_{\alpha\hat{\beta}}(x)), f_n^{\hat{\beta}}(T_{\alpha\hat{\beta}}(y))) \\ &= (0.5) d_E(T_{\alpha\hat{\beta}}(x), T_{\alpha\hat{\beta}}(y)) \\ &= (0.5) d_{\alpha\hat{\beta}}(x, y), \end{aligned}$$

for all $x, y \in \mathcal{T}$, $n = 1, 2, 3, 4$, and $\alpha \in (0,1)^3$. □

An example of $T_{\beta\alpha}$ applied to a picture is illustrated in Figure 2, for $\alpha = (0.65, 0.65, 0.65)$ and $\beta = \hat{\beta}$ where $\hat{\beta} = (0.5, 0.5, 0.5)$. The meaning of "a picture of transformation on the Euclidean plane applied to a picture" is intuitively obvious; it is discussed objectively in Section 2.2 of [5]. The picture on the left in Figure 2, a Cartesian grid masked by the triangle $\mathcal{T}$, is the "before" image, $\mathfrak{P}$, while the picture on the right is the "after" image, $T_{\hat{\beta}\alpha}(\mathfrak{P})$. Notice how straight line segments on the left are transformed into fractal paths on the right. These paths represent geodesics of $d_{\alpha\hat{\beta}}$. In fact, using the nomenclature of Kameyama [14], it can be demonstrated that $d_{\alpha\hat{\beta}} = D_{\hat{\beta}}(\mathcal{F}_\alpha|_\mathcal{T})$, the standard pseudodistance $D_{\hat{\beta}}$ with metric polyratio $\hat{\beta}$, for the topological self-similar system $\mathcal{F}_\alpha|_\mathcal{T} := (\mathcal{T}, \{f_n^\alpha\}, \pi_\alpha)$.

We notice that the shift map $S : \overline{\Omega}_\mathcal{F} \to \overline{\Omega}_\mathcal{F}$ respects the relationship $\mathcal{C} \prec \mathcal{C}$, as discussed in Remark 3.4. It follows that the dynamical system $T_\alpha : \mathcal{T} \to \mathcal{T}$, defined by $T_\alpha = \pi_\alpha \circ S \circ \tau_\alpha$, is continuous. It is readily seen that T_α maps $\mathcal{T}$ onto itself, with

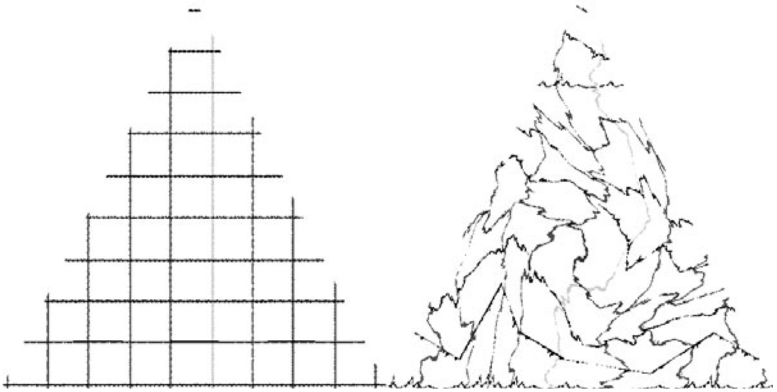

FIGURE 2. Euclidean geodesics within a triangle, on the left, are transformed by $T_{\hat{\beta}\alpha}$ into nondifferentiable paths, on the right, that are geodesics for the Kameyama metric $D_{\hat{\beta}}(\mathcal{F}_\alpha)$.

most points having four distinct preimages. The entropy of T_α is $\ln 4$, the same as that of the shift map acting on the code space of four symbols. Note however, that $T_\alpha(x)$ goes continuously clockwise three times round $\partial\mathcal{T}$ when x goes clockwise once round $\partial\mathcal{T}$. The two dynamical systems T_α, T_β are topologically conjugate, with $T_\alpha = T_{\ \beta\alpha} \circ T_\beta \circ T_{\alpha\beta}$ for all α, β. The action of T_α on some of the points of $\mathcal{T}$ is illustrated in the top left panel of Figure 3. The other panels illustrate the dynamics of the five other possible families of affine IFSs, that can be constructed similarly to $\mathcal{F}_\alpha$. Each family has a constant address structure. Thus we obtain six families of homeomorphisms on $\mathcal{T}$. Of these, only three families are distinct in the sense that no pair is conjugate via a Euclidean transformation.

6. Fractal transformations generated by bilinear functions

Let $\mathcal{R} = [0,1]^2 \subset \mathbb{R}^2$ denote the unit square, with vertices $A = (0,0), B = (1,0), C = (1,1), D = (0,1)$. Let P, Q, R, S denote, in cyclic order, the successive vertices of a possibly degenerate quadrilateral, as illustrated for example in Figure 4.

Then we uniquely define a bilinear function $\mathcal{B} : \mathcal{R} \to \mathcal{R}$ such that

$$\mathcal{B}(ABCD) = PQRS$$

by

$$\mathcal{B}(x,y) = P + x(Q-P) + y(S-P) + xy(R+P-Q-S).$$

This transformation acts affinely on any straight line that is parallel to either the x-axis or the y-axis. For example, if $\mathcal{B}|_{AB} : AB \to PQ$ is the restriction to AB of $\mathcal{B}$ and if $\mathcal{Q} : \mathbb{R}^2 \to \mathbb{R}^2$ is the affine function defined by $\mathcal{Q}(x,y) = P + x(Q-P) + y(S-P)$, then $\mathcal{Q}|_{AB} = \mathcal{B}|_{AB}$. As we illustrate, this property

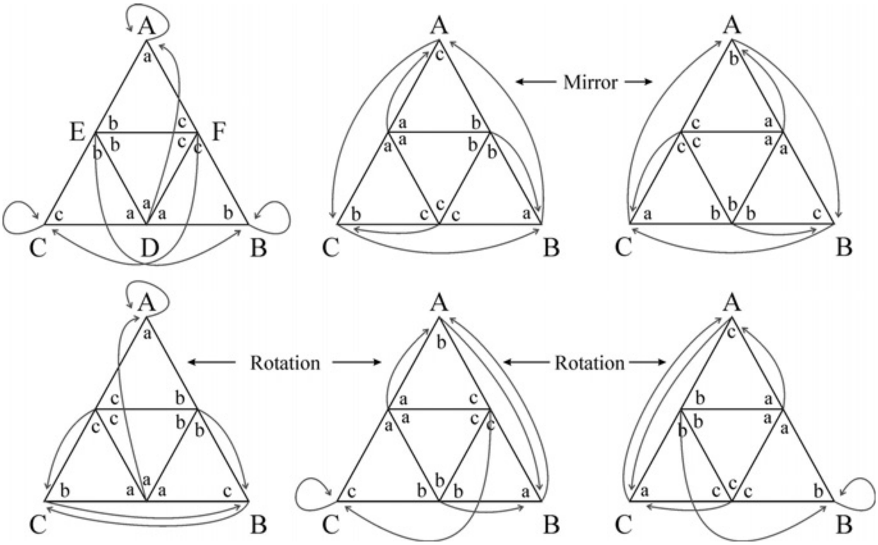

FIGURE 3. Three distinct families of fractal homeomorphisms, which are not conjugate under any affine transformation, are generated by orienting the four subtriangles $f_n(ABC) = abc$, in one of six ways. In each case the corresponding tops dynamical system is four-to-one, at almost all points, and continuous: its action on the points $A, B, \ldots, F$ is illustrated.

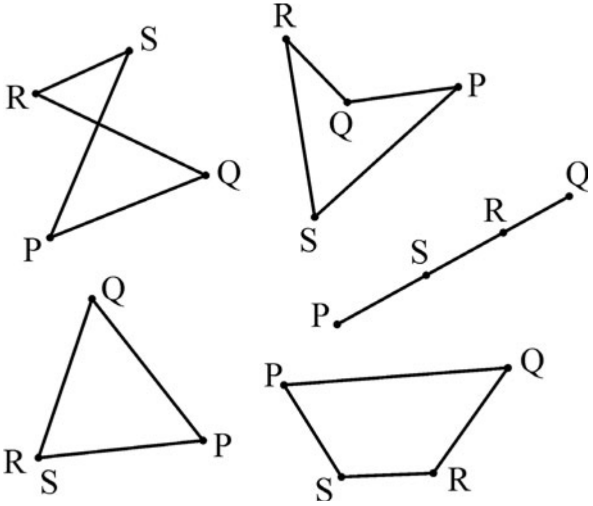

FIGURE 4. Possibly degenerate quadrilaterals with vertices P, Q, R, S in cyclic order.

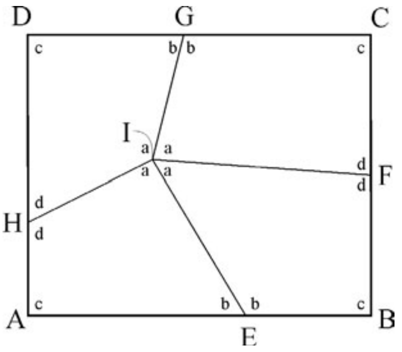

FIGURE 5. The four quadrilaterals $IEAH$, $IEBF$, $IGCF$, $IGDH$, define four bilinear transformations that are contractive with respect to an appropriately chosen metric that is Lipshitz equivalent to the Euclidean metric.

makes it easy to construct elaborate parameterized families of bilinear IFSs with constant address structures. But first we need conditions under which bilinear IFSs are point-fibred.

6.1. Contractivity of bilinear transformations

The following theorem provides practical sufficient conditions for a bilinear IFS to be hyperbolic.

Theorem 6.1. *The bilinear transformation* $\mathcal{B} : \mathcal{R} \to \mathcal{R}$ *defined by* $\mathcal{B}(x, y) = P + (Q - P)x + (S - P)y + (P - Q + R - S)xy$ *where* $P, Q, R, S \in \mathcal{R}$, *is contractive with respect to the metric* $d_{\gamma,\theta}$ *defined by* $d_{\gamma,\theta}((x_1, y_1), (x_2, y_2)) = \gamma|x_1 - x_2| + \theta|y_1 - y_2|$ *for some choice of* $\gamma, \theta > 0$ *if*

$$1 - \alpha(x, y) + \beta(x, y) > 0 \tag{6.1}$$

for all $x, y \in [0, 1]$ *where*

$$\alpha(x, y) = |(R_1 - S_1)y + (Q_1 - P_1)(1 - y)| + |(R_2 - Q_2)x + (S_2 - P_2)(1 - x)| \tag{6.2}$$

and

$$\beta(x, y) = ||((R - S)y + (Q - P)(1 - y)) \times ((R - Q)x + (S - P)(1 - x))||\,. \tag{6.3}$$

The condition (6.1) *is satisfied if*

$$\begin{aligned} &1 + 2 \min \{\text{area}(\triangle QRS), \text{area}(\triangle RSP), \text{area}(\triangle SPQ), \text{area}(\triangle PQR)\} \\ &> \max \{|R_1 - S_1|, |Q_1 - P_1|\} + \max \{|R_2 - Q_2|, |S_2 - P_2|\}\,. \end{aligned} \tag{6.4}$$

Note that $d_{\gamma,\theta}$ is a metric on $\mathbb{R}^2$ Lipshitz equivalent to the Euclidean metric provided that $\gamma > 0, \theta > 0$.

Proof. We can write

$$\mathcal{B}(x,y) = (P_1 + a_1(y)x + c_1(0)y, P_2 + a_2(0)x + c_2(x)y)$$

where $a_i(y) = (R_i - S_i)y + (Q_i - P_i)(1-y)$, and $c_i(x) = (R_i - Q_i)x + (S_i - P_i)(1-x)$, for $i = 1,2$. Thus, we seek $\gamma > 0, \theta > 0$, and $0 \leq \lambda < 1$ so that for all $(x_1, y_1), (x_2, y_2) \in \mathcal{R}$ we have

$$\begin{aligned}
&d_{\gamma,\theta}(\mathcal{B}(x_1,y_1), \mathcal{B}(x_2,y_2)) \\
&= \gamma|a_1(y_1)x_1 + c_1y_1 - a_1(y_2)x_2 - c_1y_2| + \theta|a_2x_1 + c_2(x_1)y_1 - a_2x_2 - c_2(x_2)y_2| \\
&= \gamma|a_1(y_1)(x_1 - x_2) + c_1(x_2)(y_1 - y_2)| + \theta|a_2(y_1)(x_1 - x_2) + c_2(x_2)(y_1 - y_2)| \\
&\leq (|a_1(y_1)|\gamma + |a_2(y_1)|\theta)|x_1 - x_2| + (|c_1(x_2)|\gamma + |c_2(x_2)|\theta)|y_1 - y_2| \\
&\leq \gamma|x_1 - x_2| + \theta|y_1 - y_2| = d_{\gamma,\theta}((x_1,y_1),(x_2,y_2)).
\end{aligned}$$

Hence we require that, for all $x, y \in [0,1]$,

$$|a_2(y)|\,\theta \leq (\lambda - |a_1(y)|)\,\gamma \text{ and } |c_1(x)|\,\gamma \leq (\lambda - |c_2(x)|)\,\theta.$$

This is equivalent to

$$0 \leq \lambda^2 - \lambda\alpha(x,y) + \beta(x,y) \text{ for all } x, y \in [0,1]$$

where α, β are given by (6.2) and (6.3). Hence, we can find $\lambda < 1$ provided (6.1) holds. Now note that $\beta(x,y)$ is the area of a parallelogram, two sides of which meet at $(0,0)$ and are defined by the pair of vectors $(R-S)y + (Q-P)(1-y)$ and $(R-Q)x + (S-P)(1-x)$. These vectors, in turn, are convex combinations of the two pairs of opposite sides of $PQRS$. It follows that $\beta(x,y) \geq 2\min\{\text{area}(\triangle QRS), \text{area}(\triangle RSP), \text{area}(\triangle SPQ), \text{area}(\triangle PQR)\}$. We also find

$$\alpha(x,y) \leq \max\{|R_1 - S_1|, |Q_1 - P_1|\} + \max\{|R_2 - Q_2|, |S_2 - P_2|\}.$$

Equation 6.4 follows at once. □

If P, Q, R, S are the vertices of a trapezium with sides PS and QR parallel to the y-axis, then $1 - \alpha(x,y) + \beta(x,y) = (1 - |Q_1 - P_1|)(1 - |QR|x - |PS|(1-x))$ is strictly positive for all $x, y \in [0,1]$, provided that $|Q_1 - P_1| < 1, |QR| < 1$, and $|PS| < 1$. From this it follows that the parameterized family of IFSs $\mathcal{F}_\gamma$ defined in equation (6.6) is hyperbolic. In a similar manner it is straightforward to construct other families of hyperbolic bilinear IFSs whose attractors are $\mathcal{R}$, as suggested for example by Figure 5.

An example for which Theorem 6.1 does not imply contractivity is obtained by choosing $Q = (0.2, 0.9)$, $R = (0.9, 0.1)$, $S = (0.1, 0.9)$. Then, regardless of the location of P in $\mathcal{R}$, we have $\alpha(1,1) = 1.6 > 1 + \beta(1,1) = 1.08$.

6.2. Box-counting dimensions

Let N be a positive integer. Let

$$0 = x_0 < x_1 < \cdots < x_N = 1.$$

Let $L_n : [0,1] \to [x_{n-1}, x_n]$ be the unique affine transformation, of the form $L_n(x) = a_n x + b_n$, such that $L_n(0) = x_{n-1}$ and $L_n(1) = x_n$, for $n = 1, 2, \ldots, N$. Let $0 \leq l_j \leq u_j < 1$ and $s_j = u_j - l_j$ for $j = 0, 1, \ldots, N$. Let Q_n denote the trapezium with vertices (x_{n-1}, l_{n-1}), (x_n, l_n), (x_n, u_n), and (x_{n-1}, u_{n-1}). Then we define $f_n : \mathcal{R} \to Q_n$ by

$$f_n(x, y) = (L_n(x), c_n x + [s_{n-1} + (s_n - s_{n-1})x]\, y + l_{n-1}),$$

where $c_n = l_n - l_{n-1}$. It is readily verified that each f_n is bilinear and, using Theorem 6.1, that the IFS $\mathcal{F} := (\mathcal{R}, f_1, f_2, \ldots, f_N)$ is hyperbolic. Using standard methods, [3], it is readily verified that the attractor $A_{\mathcal{F}}$ of $\mathcal{F}$ is the graph $\Gamma(g)$ of a continuous function $g : [0,1] \to [0,1]$.

For present purposes we define the box-counting dimension of $A_{\mathcal{F}}$ to be

$$\dim A_{\mathcal{F}} := \lim_{\varepsilon \to 0+} \frac{\log \mathcal{N}_{\varepsilon}(A_{\mathcal{F}})}{\log \varepsilon^{-1}} \tag{6.5}$$

where $\mathcal{N}_{\varepsilon}(A_{\mathcal{F}})$ is the minimum number of square boxes, with sides parallel to the axes, whose union contains $A_{\mathcal{F}}$. By the statement "$\dim A_{\mathcal{F}} = D$" we mean that the limit in equation (6.5) exists and equals D.

Theorem 6.2. [8] *Let $\mathcal{F}$ denote the bilinear IFS defined above, and let $A_{\mathcal{F}}$ denote its attractor. Let $a_n = 1/N$ for $n = 1, 2, \ldots, N$ and let $\sum_{n=1}^{N} \frac{s_{n-1}+s_n}{2} > 1$. If $A_{\mathcal{F}}$ is not a straight line segment then*

$$\dim A_{\mathcal{F}} = 1 + \frac{\log \sum_{n=1}^{N} \frac{s_{n-1}+s_n}{2}}{\log N}$$

Information about $\dim A_{\mathcal{F}} = \dim \Gamma(g)$ provides information about the smoothness of g because $\dim \Gamma(g)$ is related to Hölder exponents associated with g, see [25], Section 12.5, for example.

6.3. A family of fractal homeomorphisms generated by bilinear transformations

The following example is considered in [8], which provides more details and the proofs of results stated here.

Let $I = (0, p)$, $J = (0, q)$, $F = (1, q)$, $G = (1, p)$, $E = (0.5, 0)$, $K = (0.5, r)$, $L = (0.5, s)$ where $0 < q < p < 1$ and $0 < s < r < 1$, as illustrated in Figure 6. Define bilinear functions $\mathcal{B}_n : \mathcal{R} \to \mathcal{R}$ for $n = 1, \ldots, 6$ by

$$\mathcal{B}_1(ABCD) = AELJ, \mathcal{B}_2(ABCD) = EBFL, \mathcal{B}_3(ABCD) = JLKI,$$
$$\mathcal{B}_4(ABCD) = LFGK, \mathcal{B}_5(ABCD) = IKHD, \mathcal{B}_6(ABCD) = KGCH.$$

Define a family of iterated function systems, dependent on the vector of parameters $\gamma = (p, q, r, s)$, by

$$\mathcal{F}_{\gamma} = (\mathcal{R}, \{\mathcal{B}_n\}_{n=1}^{6}, \pi_{\gamma}). \tag{6.6}$$

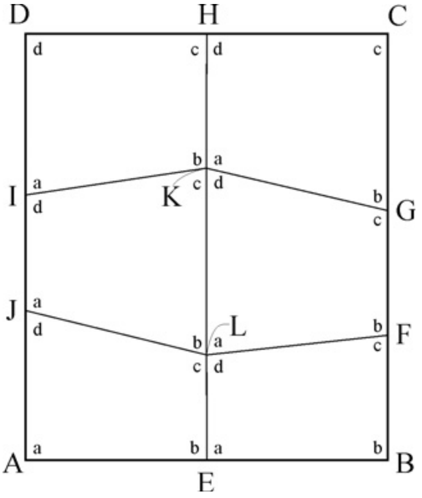

Figure 6. The arrangement of quadrilaterals used in in Section 6.3. The letters a, b, c,and d in the corners of a quadrilateral indicate the images of the points A, B, C, and D, respectively, under the corresponding bilinear transformation.

Then Theorem 6.1 provides that, for all admissible γ, $\mathcal{F}_\gamma$ is hyperbolic, with attractor $A = \pi_\gamma(\Omega) = \mathcal{R}$. It is also straightforward to show, using the affinity of each $\mathcal{B}_n$ when restricted to any side of $\mathcal{R}$, that the address structure $\mathcal{C}_\gamma = \mathcal{C}$ is independent of γ. It follows that we can define a family of fractal homeomorphisms $T_{\gamma\delta} : \mathcal{R} \to \mathcal{R}$ by $T_{\gamma\delta} = \pi_\delta \circ \tau_\gamma$ for all admissible γ, δ. We remark, however, that the shift operator $S : \Omega_\gamma \to \Omega_\gamma$ does not respect the address structure $\mathcal{C}_\gamma$ and consequently the tops dynamical system $T_\gamma := \pi_\gamma \circ S \circ \tau_\gamma : \mathcal{R} \to \mathcal{R}$ is not continuous, in contrast to the examples in Section 5.

Similarly to the transformations in Section 5.2, we have

$$T_{\epsilon\gamma} \circ T_{\epsilon\delta}^{-1} = T_{\gamma\delta}$$

for all admissible γ, δ, ϵ. In particular, we can obtain information about the structure and smoothness of $T_{\gamma\delta}$ by studying $T_{\hat{\epsilon}\gamma} : \mathcal{R} \to \mathcal{R}$, for all admissible γ, in the case where $\hat{\epsilon}$ denotes the parameter set $p = r = 2/3$, $q = s = 1/3$.

We observe that $T_{\gamma\delta}(L_x) = L_x$ for all $x \in [0, 1]$ where L_x is the line segment $\{(x, y) : 0 \leq y \leq 1\}$. Consequently, if $\Gamma(f) \subset \mathcal{R}$ denotes the graph of a continuous function $f : [0, 1] \to [0, 1]$, then $T_{\gamma\delta}(\Gamma(f))$ is also the graph of a continuous function from $[0, 1]$ to itself. So let $C[0, 1]$ denote the set of continuous functions from $[0, 1]$ into itself, with metric $d_{C[0,1]}(f, g) = \max\{|f(x) - g(x)| : x \in [0, 1]\}$. Then $T_{\gamma\delta} : \mathcal{R} \to \mathcal{R}$ induces a continuous transformation $\tilde{T}_{\gamma\delta} : C[0, 1] \to C[0, 1]$, defined by $\tilde{T}_{\gamma\delta}(f) = g$ where $g \in C[0, 1]$ is uniquely defined by $T_{\gamma\delta}(\Gamma(f)) = \Gamma(g)$.

Let $f_c \in C[0, 1]$ be defined by $f_c(x) = c$ for $c \in [0, 1]$. Information about the smoothness of $T_{\hat{\epsilon}\gamma}$ is obtained by looking at the functions $g_c := \tilde{T}_{\hat{\epsilon}\gamma}(f_c)$ for various values of c. In [8] it is proved that

$$g_{c_1}(x) < g_{c_2}(x) \text{ whenever } 0 \leq c_1 < c_2 \leq 1,$$

for all $x \in [0,1]$ and all admissible γ. Since $\mathcal{R} = \cup \{T_{\gamma\delta}(\Gamma(f_c)) : c \in [0,1]\}$, for each admissible γ, δ, it now follows that the graphs of the set of functions $\{g_c : c \in [0,1]\}$ tile $\mathcal{R}$, for each admissible γ. For example, when $\gamma = \hat{\epsilon}$, we have $g_c = f_c$ and the graphs of the set of functions $\{f_c : c \in [0,1]\}$ tile $\mathcal{R}$.

In [8] it is proved that

$$f_0 = g_0 < g_{1/2} < g_1 = f_1,$$

where $\Gamma(g_0)$ is the attractor of the IFS $\mathcal{F}_\gamma^{(1)} := (\mathcal{R}, \mathcal{B}_1, \mathcal{B}_2)$, $\Gamma\left(g_{1/2}\right)$ is the attractor of the IFS $\mathcal{F}_\gamma^{(2)} := (\mathcal{R}, \mathcal{B}_3, \mathcal{B}_4)$, and $\Gamma(g_2)$ is the attractor of the IFS $\mathcal{F}_\gamma^{(3)} := (\mathcal{R}, \mathcal{B}_5, \mathcal{B}_6)$. Furthermore, by Theorem 6.2, if $\Gamma\left(g_{\mathbf{1/2}}\right)$ is not a line segment and $(p - q + r - s) > 1$ then

$$\dim\Gamma(g_0) = 1, \dim\Gamma\left(g_{1/2}\right) = 1 + \frac{\log\,(p-q+r-s)}{\log 2}, \dim\Gamma(g_1) = 1.$$

So for example if $p = 5/8$, $q = 1/8$, $r = 7/8$, $s = 2/8$ then $\dim\Gamma\left(g_{\mathbf{1/2}}\right) = (\log 9 - 2\log 2)/\log 2 = 1.1699\ldots$. So the image under $T_{\hat{\epsilon}\gamma}$ of the three line segments $\Gamma(f_0)$, $\Gamma\left(f_{1/2}\right)$, $\Gamma(f_1)$ is a sandwich of three curves, the upper and lower having dimension one and the middle curve having box-counting dimension greater than one and less than two. This sandwich is repeated at finer and finer scales, as can be seen by applying compositions of finite sequences of operators from the set $\left\{\mathcal{F}_\gamma^{(1)}, \mathcal{F}_\gamma^{(2)}, \mathcal{F}_\gamma^{(3)}\right\}$ to the sandwich. This notion is implicit in Figure 7.

Acknowledgements

I thank Louisa Barnsley for the illustrations. I thank Maria Moszyńska for helpful discussions concerning Theorem 4.2. I thank Üta Freiberg for interesting observations about point-fibred affine IFSs on triangles, at an early stage of the work reported here. I also thank David C. Wilson, Peter Massopust, and Andrew Vince, for observations and comments related to aspects of this work.

References

[1] Ross Atkins, M.F. Barnsley, David C. Wilson, Andrew Vince, A characterization of point-fibred affine iterated function systems, *Preprint*, submitted for publication, (2009).

[2] M.F. Barnsley and S.G. Demko, Iterated function systems and the global construction of fractals, *Proc. Roy. Soc. London Ser. A* **399** (1985) 243–275.

[3] M.F. Barnsley, Fractal functions and interpolation. *Constr. Approx.* **2** (1986), no. 4, 303–329.

[4] ———, Theory and application of fractal tops. 3–20, *Fractals in Engineering: New Trends in Theory and Applications.* Lévy-Véhel J.; Lutton, E. (eds.) Springer-Verlag, London Limited, 2005.

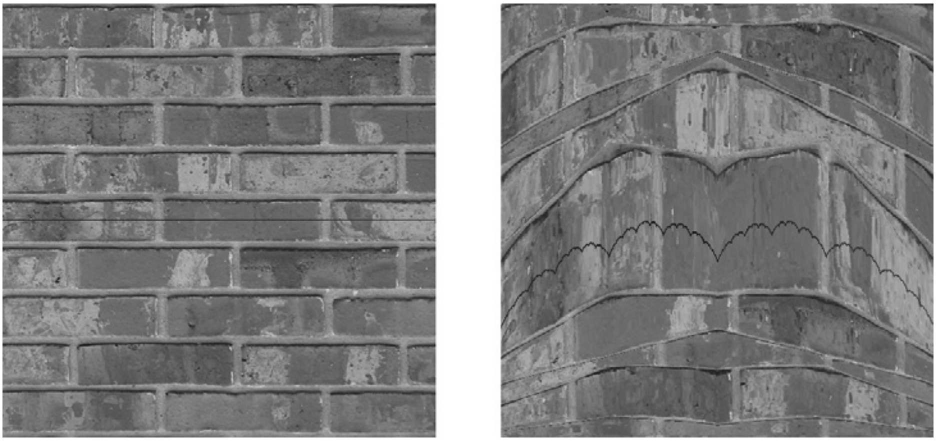

FIGURE 7. The image on the left, which is supported on $\mathcal{R}$, is transformed to become the image on the right under the fractal homeomorphism $T_{\hat{\epsilon}\gamma}$ discussed at the end of Section 6.3. Horizontal lines on the left are transformed to become the graphs of fractal interpolation functions. For example the horizontal line through the center of the image on the left becomes a curve with fractal dimension 1.1699..., illustrated in black in the image on the right.

[5] ———, *Superfractals*, Cambridge University Press, Cambridge, 2006.

[6] ———, Transformations between self-referential sets, *Amer. Math. Monthly* **116** (2009) 291–304.

[7] M.F. Barnsley, J.E. Hutchinson, New methods in fractal imaging, *Proceedings of the International Conference on Computer Graphics, Imaging and Visualization,* (*July* 26–28*,* 2006), IEEE Society, Washington D.C. 296–301.

[8] M.F. Barnsley, P. Massopust, M. Porter, Fractal interpolation and superfractals using bilinear transformations, Preprint, 2009.

[9] Leonard M. Blumenthal, *Theory and Applications of Distance Geometry*, Chelsea Publishing Company, New York, 1970.

[10] Herbert Buseman, *The Geometry of Geodesics*, Academic Press, New York, 1955.

[11] August Florian, On a metric for a class of compact convex sets, *Geometriae Dedicata* **30** (1989) 69–80.

[12] M. Hata, On the structure of self-similar sets, *Japan J. Appl. Math.* **2** (1985) 381–414.

[13] J.E. Hutchinson, Fractals and self-similarity*, Indiana Univ. Math. J.* **30** (1981) 713–747.

[14] Atsushi Kameyama, Distances on topological self-similar sets*, Proceedings of Symposia in Pure Mathematics*, Volume **72.1**, 2004.

[15] ———, Self-similar sets from the topological point of view*, Japan J. Ind. Appl. Math.* **10** (1993) 85–95.

[16] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems. With a Supplementary Chapter by Katok and Leonardo Mendoza,* Cambridge University Press, Cambridge, 1995.

[17] Bernd Kieninger, *Iterated Function Systems on Compact Hausdorff Spaces,* Shaker Verlag, Aachen, 2002.

[18] J. Kigami, Harmonic calculus on p.c.f. self-similar sets, *Trans. Amer. Math. Soc.* **335** (1993) 721–755.

[19] J. Kigami, *Analysis on Fractals,* Cambridge University Press, Cambridge, 2001.

[20] Maria Moszyńska, *Selected Topics in Convex Geometry,* Birkhäuser, Boston, 2006.

[21] James R. Munkres, *Topology*, Prentice Hall, 2000.

[22] R. Tyrrel Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[23] R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory,* Cambridge University Press, 1993.

[24] A.C. Thompson, *Minkowski Geometry,* Cambridge University Press, 1996.

[25] Claude Tricot, *Curves and Fractal Dimension*, Springer-Verlag, Berlin, 1995.

[26] R.F. Williams, Composition of contractions*, Bol. da Soc. Brasil de Mat.* **2** (1971) 55–59.

[27] Roger Webster, *Convexity*, Oxford University Press, Oxford, 1994.

Michael F. Barnsley
Department of Mathematics
Australian National University
Canberra, ACT, Australia
e-mail: michael.barnsley@maths.anu.edu.au
mbarnsley@aol.com
URL: http://www.superfractals.com

Progress in Probability, Vol. 61, 251–268

Geometric Realizations of Hyperbolic Unimodular Substitutions

Maki Furukado, Shunji Ito and Hui Rao

Abstract. We generalize the construction of Rauzy fractals to hyperbolic substitutions. A fractal-domain-exchange transformation is defined, and it is proved that this transformation is measure-theoretically isomorphic to the substitution dynamical system.

Mathematics Subject Classification (2000). Primary 37B10; Secondary 37D20.

Keywords. Hyperbolic substitution, Rauzy fractal, substitution dynamical system.

1. Introduction

1.1. Substitution dynamical systems

First we introduce some notations on substitutions. Let $\mathcal{A} = \{1, \ldots, d\}$ be an alphabet, $d \geq 2$. Let $\mathcal{A}^* = \cup_{n \geq 0} \mathcal{A}^n$ be the set of finite words. A substitution is a function $\sigma : \mathcal{A} \mapsto \mathcal{A}^*$. The incidence matrix of σ is $M_\sigma = M = (m_{ij})_{1 \leq i,j \leq d}$, where m_{ij} is the number of occurrences of i in $\sigma(j)$. A matrix A is *primitive* if there exists a positive integer N such that A^N is a positive matrix. We will always assume that the incidence matrix M is primitive.

An infinite word $s \in \mathcal{A}^\infty$ is a *fixed point* of σ if $\sigma(s) = s$; it is a *periodic point* if $\sigma^k(s) = s$ for some integer $k \geq 1$. A primitive substitution has at least one periodic point.

Let T be the (left)-shift operator on the symbolic space $\mathcal{A}^\infty$, let

$$\Omega = \overline{\{T^n(s);\ n \geq 0\}}$$

be the orbit closure of the fixed point s of σ. If the substitution is primitive, it is well known that $\{\Omega, T\}$ is uniquely ergodic [25, 29]. We denote the unique ergodic probability measure by ν, and the dynamical system $\{\Omega, T, \nu\}$ is called the *substitution dynamical system* of σ.

The third author is supported by CNSF no. 10631040.

Substitution dynamical systems form an important class of symbolic dynamics. Spectral analysis and geometrical realizations of substitution dynamics are two interesting problems. (We are particularly interested in when the spectrum is discrete.) These two problems are closely related to each other by the famous representation theorem: *An ergodic measure-preserving transformation has discrete spectrum if and only if it is conjugate to an ergodic rotation on some compact abelian group* (cf. [35] pp. 73).

Substitutions of constant length. A substitution σ is said to be of constant length if $|\sigma(1)| = |\sigma(2)| = \cdots = |\sigma(d)|$. Dekking [9] introduced a notion of coincidence for constant length substitution and proved that σ has discrete spectrum if and only if σ admits a coincidence.

Substitutions of Pisot type. Another interesting class of substitutions are Pisot substitutions. A substitution σ is said to be *a Pisot substitution* if the Perron-Frobenius eigenvalue of M_σ is a Pisot number (an algebraic integer whose algebraic conjugates have modulus strictly less than 1). The following conjecture plays a central role.

Pisot spectral conjecture: Any Pisot substitution has discrete spectrum.

This problem is also related to a geometrical realization introduced by Rauzy [30], which is called *Rauzy fractals* or *atomic surfaces.* For example, let τ denote the Fibonacci substitution: $1 \mapsto 12, 2 \mapsto 1$. Then the Rauzy fractals of τ are intervals, and the substitution dynamical systems of τ is metrically conjugate to a two-interval-exchange, which can be regarded as a rotation of one-dimensional torus. Hence τ has discrete spectrum by the representation theorem.

Let η denote the Rauzy substitution

$$1 \mapsto 12,\ 2 \mapsto 13,\ 3 \mapsto 1.$$

Its Rauzy fractal consists of three piece X_1, X_2, X_3 and a domain-exchange transformation D can be defined on $X = X_1 \cup X_2 \cup X_3$. (See Figure 1.) Since X admits a lattice tiling of $\mathbb{R}^2$, D can be regarded as a rotation of two-dimensional torus. Rauzy showed that the substitution dynamical system of η is metrically conjugate to D and hence has discrete spectrum.

The Pisot spectral conjecture is confirmed for two-letter substitutions in [6] and it is widely open for many-letter substitutions.

Rauzy fractals of Pisot substitutions have received extensive studies in the past 25 years ([23, 1, 33, 8, 32, 2, 3, 6, 28, 7] etc). It is an interesting object in number theory, tiling theory, spectral theory and dynamical systems.

Hyperbolic substitutions. It is natural to generalize Rauzy's construction to non-Pisot substitutions. Early work in this direction was done by Holton and Zamboni [15]. They constructed *complex geometric realizations* of a large class of substitutions. However, as we shall see, the dynamical property of this realization is not satisfactory and it does not give us a conjugacy of the substitution dynamical system.

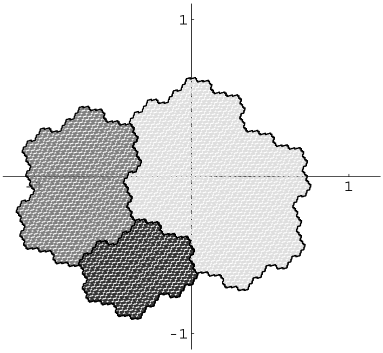

FIGURE 1. Rauzy fractals $\bigcup_{i=1,2,3} X_i$ of Rauzy substitution.

The goal of the present paper is to construct a geometrical realization of unimodular hyperbolic substitutions in the irreducible case. Precisely, we construct a *fractal domain-exchange transformation* which is isomorphic to the substitution dynamical system.

1.2. Rauzy fractals of hyperbolic substitutions

A substitution σ is *unimodular* if $\det M = \pm 1$; it is *of irreducible type* if the characteristic polynomial of M is irreducible over $\mathbb{Q}$. A substitution is *hyperbolic* if there are no eigenvalues of the incidence matrix on the unit circle.

In this paper, we only consider unimodular hyperbolic substitutions of irreducible type. Let σ be such a substitution. Then the eigenvalues $\lambda_1, \ldots, \lambda_d$ of the incidence matrix M satisfy

$$|\lambda_1|, \ldots, |\lambda_m| < 1 < |\lambda_{m+1}|, \ldots, |\lambda_d|,$$

and the linear map of $\mathbb{R}^d$ defined by M has a stable space P of dimension m and an unstable space V of dimension $d-m$. According to the direct sum $\mathbb{R}^d = V \oplus P$, we define two natural projections

$$\pi\colon\ \mathbb{R}^d \mapsto P, \quad \pi'\colon\ \mathbb{R}^d \mapsto V.$$

Since the Perron-Frobenius eigenvector is non-rational, the projections of the lattice $\mathbb{Z}^d$ is dense in P as well as V; it is also one-to-one on $\mathbb{Z}^d$ from $\mathbb{Z}^d$ to P as well as V : for any $x, y \in \mathbb{Z}^d$, $\pi(x) = \pi(y)$ (or $\pi'(x) = \pi'(y)$) if and only if $x = y$. Denote by $\vec{e}_1, \ldots, \vec{e}_d$ the canonical basis of $\mathbb{R}^d$. Define a map $f:\ \mathcal{A}^* \to \mathbb{Z}^d$ by

(i) $f(\epsilon) = 0$, where ϵ denotes the empty word;
(ii) $f(i) = \vec{e}_i$ for $1 \le i \le d$ and $f(UV) = f(U) + f(V)$ for $U, V \in \mathcal{A}^*$.

The following construction is an analogue of that of Rauzy fractals. Let $s = s_1 s_2 \ldots$ be a periodic point of σ. Set $Y = \{y_k = f(s_1 \ldots s_{k-1}) :\ k \ge 1\}$. Then Y is a subset of $\mathbb{Z}^d$. If we connect every two consecutive points by a line segment, we obtain a broken line from the origin to infinity. The *Rauzy fractal* of σ is the closure $X = \overline{\pi(Y)}$ of the projection of Y on the stable space P. Furthermore, let

$$Y_i = \{f(s_1 \dots s_{k-1}) : \ s_k = i, \ k \geq 1\} \text{ and } X_i = \overline{\pi(Y_i)}$$

for $1 \leq i \leq d$. We call the family $\{X_i\}_{1\leq i\leq d}$ *the partial Rauzy fractals* of σ. Clearly $X = \cup_{i=1}^{d} X_i$.

Example. Consider the substitution:

$$1 \to 14, \ 2 \to 3, \ 3 \to 423, \ 4 \to 142.$$

The incidence matrix is $M = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$, the characteristic polynomial of M is $\Phi_M(x) = x^4 - 3x^3 + x^2 + x + 1$, and the eigenvalues are $\{2.28879, \ 1.38939, -0.339093 \pm 0.44663i\}$. Since the eigenvalues inside the unit circle are complex conjugates, the contraction matrix of the Rauzy fractals is a similitude. (See Figure 2.)

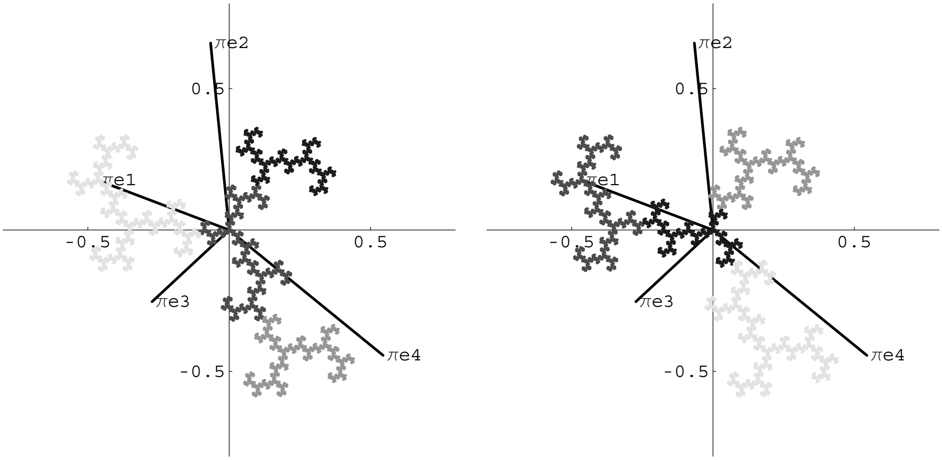

FIGURE 2. X_i and $X'_i = X_i + \pi e_i$.

1.3. Self-similarity and open set condition

First we show that the partial Rauzy fractals X_j have a self-similar structure, namely, MX_j is a union of translation copies of $X_1, \dots, X_d$. We also show that the unions are not overlapping; in terminology of IFS theory, the graph IFS of the partial Rauzy fractals satisfy an *open set condition.* (For Graph-IFS and open set condition, we refer to [27] and [12].)

Let us denote $\sigma(i) = a_{i1} \dots a_{il_i}$ and set

$$\mathcal{D}_{ij} = \{M^{-1}\pi \circ f(a_{j1} \dots a_{j(k-1)});\ a_{jk} = i\}.$$

Theorem 1.1. *Let σ be an unimodular hyperbolic substitution of irreducible type. Then the partial atomic surfaces $\{X_i\}_{i=1}^d$ are compact and satisfy the set equations*

$$M^{-1}X_i = \bigcup_{j=1}^{d} (X_j + \mathcal{D}_{ij})\,. \tag{1.1}$$

Since the compact sets satisfying (1.1) are unique [27, 12], the atomic surfaces do not depend on the choice of the periodic point s.

If σ is a Pisot substitution, X_i have non-empty interior and usually have fractal boundary. (See Figure 1.) In case that the substitution is not Pisot, the Rauzy fractals have empty interiors and they are indeed fractal sets. (See Figure 2.)

Theorem 1.2. *Let σ be a unimodular hyperbolic substitution of irreducible type. Then the graph-IFS* (1.1) *satisfies the open set condition.*

We note that the construction of [15] usually does not satisfy the open set condition.

The validity of the open set condition in the hyperbolic case is not as obvious as in the Pisot case, because the contraction matrix of the system need not be a similitude, so that the similarity dimension of the system is less than the dimension of the space. However, in a recent paper, He and Lau generalized the results of Schief [31] to some self-affine iterated function systems, and they established an algebraic criterion of the open set condition for such systems [14]. Luo and Yang [22] further generalize this result to the graph IFS case, which is crucial in our argument.

1.4. Geometric realization of substitution dynamic

For Pisot substitutions, a *domain-exchange transformation* $D : X \mapsto X$ has been defined by Rauzy [30], Arnoux and Ito [1]. We can generalize this definition to hyperbolic substitutions:

$$D(x) = x + \pi(e_k), \qquad \text{if } x \in X_k \text{ and } x \notin \bigcup_{j=1}^{k-1} X_j.$$

To show this *fractal domain-exchange transformation* (X, D) is a realization of the substitution dynamical system, we need to find a measure μ supported on X which is 'translation invariant' provided the translation is inside X, and that $X = \bigcup_{i=1}^{d} X_i$ is a disjoint union in measure μ.

The existence of a 'translation invariant' measure is confirmed by [22]. They showed that the Hausdorff measure of X w.r.t. a pseudo norm is positive and finite, *i.e.*, $0 < \mathcal{H}_w^\alpha(X) < +\infty$, where α is the Hausdorff dimension of X w.r.t. the weak norm. Hence the restriction of $\mathcal{H}_w^\alpha$ on X is such a measure. Let us denote this measure by μ.

To show that $X = \bigcup_{i=1}^{d} X_i$ is disjoint in measure μ, we need to assume that σ satisfies a *strong coincidence condition,* which was first introduced by Dekking for constant length substitution and then generalized to general substitution by Host and Levishtiz.

Theorem 1.3. *If σ satisfies the strong coincidence condition, then (X, D, μ) is measure-preserving and it is isomorphic to the dynamical system (Ω, T, ν).*

1.5. Questions

In the Pisot case, if we assume that σ satisfies a super-coincidence condition, then it is shown that (X, D, μ) is a rotation on the torus, and hence the substitution dynamical system has purely discrete spectrum [17]. However, in the non-Pisot case, we do not know how this geometrical realization is related to the spectral property of the substitution dynamical system.

The paper is organized as follows. In Section 2, we recall some recent results on single-matrix graph IFS. In Section 3, we discuss the self-similar structure of atomic surfaces and prove the open set condition holds. In Section 4, we construct a fractal domain-exchange transformation (X, D, μ), and we show that (X, D, μ) is isomorphic to (Ω, T, ν) in Section 5.

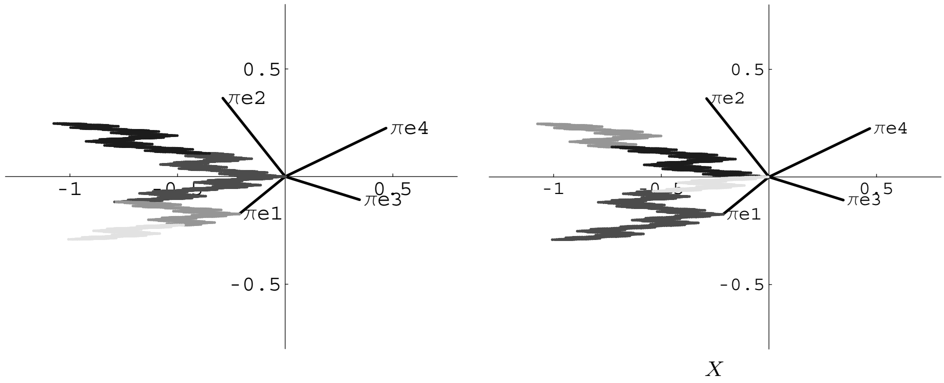

FIGURE 3. X_i and X_i' of $\sigma : \ 1 \to 123, \ 2 \to 1, \ 3 \to 14, \ 4 \to 3$. The eigenvalues are $\{2.09529, \ -1.35567, \ 0.73764, \ -0.47726\}$.

2. Single-matrix graph IFS

2.1. Graph IFS

Let (V, Γ) be a directed graph with vertex set $V = \{1, \ldots, N\}$ and edge set Γ. We call $\{f_e; \ e \in \Gamma\}$, a collection of contractions $f_e : \mathbb{R}^d \mapsto \mathbb{R}^d$, a *graph-directed iterated function system* (graph IFS).

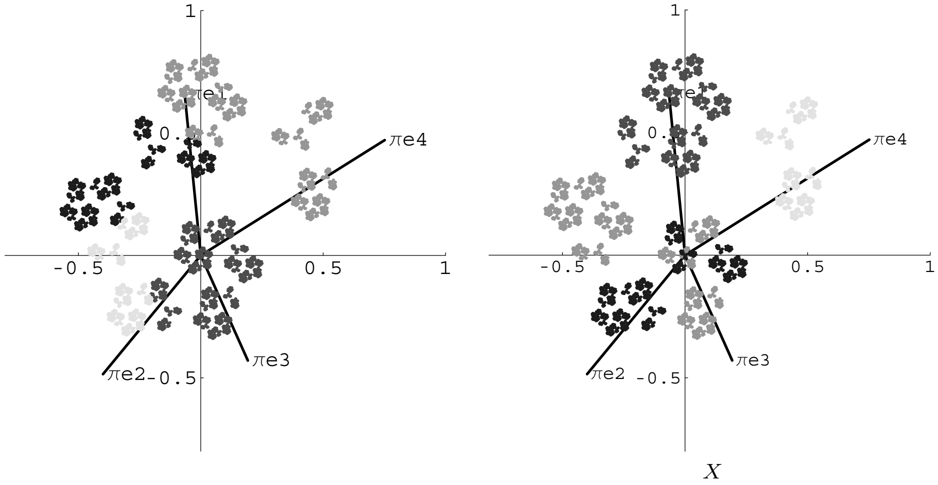

FIGURE 4. X_i and $X_i' = X_i + \pi e_i$ of $1 \to 12,\ 2 \to 3,\ 3 \to 42,\ 4 \to 1$. The eigenvalues are $\{1.5129,\ -1.1787,\ 0.33292 \pm 0.67077i\}$.

Let Γ_{ij} be the set of edges from vertex i to j, then there are unique non-empty compact sets $\{E_i\}_{i=1}^N$ satisfying ([27])

$$E_i = \bigcup_{j=1}^{N} \bigcup_{e \in \Gamma_{ij}} f_e(E_j), \quad 1 \le i \le N. \tag{2.1}$$

We call $(E_1, \ldots, E_N)$ the *invariant sets* of the graph IFS. (See also the book of Falconer [12].)

The graph IFS is said to satisfy the *open set condition* (OSC), if there exist open sets $U_1, \ldots, U_N$ such that

$$\bigcup_{j=1}^{N} \bigcup_{e \in \Gamma_{ij}} f_e(U_j) \subset U_i, \quad 1 \le i \le N,$$

and the left-hand side are non-overlapping unions.

Let us define $M = (m_{ij})_{1 \le i,j \le N}$ to be the associated matrix of Γ, that is, $m_{ij} = \#\Gamma_{ji}$ counts the number of edges from j to i. We say Γ is *strongly connected* if M is a primitive matrix. From now on, we will always assume that the graph under consideration is strongly connected.

2.2. Single-matrix graph IFS

We are particularly interested in graph IFS $\{f_e;\ e \in \Gamma\}$ satisfying

$$f_e = A^{-1}(x + d_e) \text{ for } e \in \Gamma, \tag{2.2}$$

where A is an expanding matrix in $M_d(\mathbb{R})$. (A matrix is *expanding* if all its eigenvalues have modulus greater than 1.) We shall call (2.2) a *single-matrix graph IFS.*

Example.

(i) McMullen's carpets [24], a class of typical self-affine IFS, are single-matrix IFS.
(ii) Self-affine tilings studied in [4, 18, 20] are single-matrix IFS.
(iii) Rauzy fractals of hyperbolic unimodular substitutions are single-matrix graph IFS.

Denote by Γ^n_{ij} the paths from vertex i to vertex j with length n. For $I = e_1 \dots e_n \in \Gamma^n_{ij}$, set $f_I(x) := f_{e_1} \circ f_{e_2} \circ \cdots f_{e_n}(x)$ and define

$$d_I := A^{n-1}d_{e_1} + A^{n-2}d_{e_2} + \cdots + Ad_{e_{n-1}} + d_{e_n},$$

then $f_I(x)$ has the form: $f_I(x) = A^{-n}(x + d_I)$. Set

$$\mathcal{D}^n_{ij} := \{d_I;\ I \in \Gamma^n_{ij}\}. \tag{2.3}$$

In this case, formula (2.1) can be simplified to $E_i = \bigcup_{j=1}^N A^{-1}(E_j + \mathcal{D}_{ij})$, which is

$$AE_i = \bigcup_{j=1}^{N} (E_j + \mathcal{D}_{ij})\,. \tag{2.4}$$

Iterating(2.4), we obtain

$$A^n E_i = \bigcup_{j=1}^{N} \left(E_j + \mathcal{D}^n_{ij}\right).$$

2.3. Pseudo-norm

Let A be a $d \times d$ real expanding matrix with $|\det A| = q$. [19] defines a pseudo-norm $w = w_A$ on $\mathbb{R}^d$ as follows (see also [14]).

Denote $B(x, r)$ the open ball with center x and radius r. Then $V = A(B(0, 1)) \backslash B(0, 1)$ is an annular region. Choose any $0 < \delta < \frac{1}{2}$ and any C^∞ function $\phi_\delta(x)$ with support in $B(0, \delta)$ such that $\phi_\delta(x) = \phi_\delta(-x)$ and $\int \phi_\delta(x)dx = 1$, define a pseudo-norm $w(x)$ in $\mathbb{R}^d$ by

$$w(x) = \sum_{n \in \mathbb{Z}} q^{-n/d} h(A^n x), \tag{2.5}$$

where $h(x) = \chi_V * \phi_\delta(x)$ is the convolution of the indicator function χ_V and $\phi_\delta(x)$.

It is shown that $w(Ax) = q^{1/d} w(x)$ for all $x \in \mathbb{R}^d$. Hence the matrix A is a similitude in this weak norm. This fact plays a central role in [14] and [22].

Let E be a subset of $\mathbb{R}^d$. Define $\operatorname{diam}_\omega E = \sup\{\omega(x - y) : x, y \in E\}$ be the ω-diameter of E. Now we can define a Hausdorff measure of E with respect to the pseudo-norm $w(x)$.

$$\mathcal{H}^{\alpha}_{w,\delta}(E) = \inf\{\sum_{i=1}^{\infty}(\mathrm{diam}_w E_i)^{\alpha} : E \subset \bigcup_i E_i, \mathrm{diam}_w E_i \leq \delta\}$$

Since $\mathcal{H}^{\alpha}_{w,\delta}(E)$ is increasing when δ tends to 0, we can define

$$\mathcal{H}^{\alpha}_{w}(E) = \lim_{\delta\to 0}\mathcal{H}^{\alpha}_{w,\delta}(E).$$

It is shown that $\mathcal{H}^{\alpha}_{w}(E)$ is an outer measure and is a regular measure on the family of Borel subsets on $\mathbb{R}^d$. It is translation invariant and has the scaling property; precisely,

$$\mathcal{H}^{\alpha}_{w}(E+x) = \mathcal{H}^{\alpha}_{w}(E) \quad \text{and} \quad \mathcal{H}^{\alpha}_{w}(A^{-1}E) = q^{-\alpha/d}\mathcal{H}^{\alpha}_{w}(E). \tag{2.6}$$

A Hausdorff dimension with respect to the pseudo-norm thus can be defined to be

$$\dim_w E = \inf\{\alpha : \mathcal{H}^{\alpha}_{w}(E) = 0\} = \sup\{\alpha : \mathcal{H}^{\alpha}_{w}(E) = \infty\}.$$

The relation of $\dim_w E$ and the classical Hausdorff dimension $\dim_H(E)$ has been studied in [14].

When A is a similitude, then $\mathcal{H}^{t}_{w}$ and $\dim_w$ become the normal Hausdorff measure and dimension respectively.

2.4. Algebraic criterion of OSC

The following criterion shows that a certain discreteness implies the open set condition.

Theorem 2.1 ([22]). *Graph IFS* (2.2) *satisfies the OSC if and only if* $\#\mathcal{D}^n_{ij} = \#\Gamma^n_{ij}$ *and there is a* $r > 0$ *such that* $\mathcal{D}^n_{ij}$ *is r-uniformly discrete for all* $1 \leq i, j \leq N$ *and* $n \geq 1$.

The above result is well known when A is a similitude or the system in consideration forms a tiling system ([20, 21]). In the IFS case, the result is proved by He and Lau by using a pseudo-norm w on $\mathbb{R}^d$.

Under the pseudo-norm w, A becomes a "similitude". Then using the method developed by Bandt and Graf [5], and Schief [31], He and Lau recovered the results of Schief. Later it is generalized to Graph-IFS by Luo and Yang [22].

Let λ be the maximal eigenvalue of M, the associate matrix of Γ.

Theorem 2.2 ([22]). *For graph IFS* (2.2), *if OSC holds, then for any* $1 \leq i \leq N$,

(i) $\alpha = \dim_{\omega} E_i = d\log\lambda/\log q$.

(ii) $0 < \mathcal{H}^{\alpha}_{\omega}(E_i) < +\infty$.

(iii) *The right-hand side of* (2.1) *is a disjoint union in measure* $\mathcal{H}^{\alpha}_{w}$.

3. Verifying the open set condition for Rauzy fractal

3.1. Notations

For $x \in \mathbb{Z}^d$, we denote by $(x, i) := \{x + \theta e_i;\ \theta \in [0,1]\}$ the segment from x to $x + e_i$. Define $y + (x, i) := (x + y, i)$. Let $\mathcal{G} = \{(x, i) :\ x \in \mathbb{Z}^d, 1 \leq i \leq d\}$. The term *segment* will refer to any element of $\mathcal{G}$.

Recall that $\sigma(i) = a_{i1} \dots a_{il_i}$.

The *inflation and substitution map* F_σ is defined as follows on the family of subsets of $\mathcal{G}$. Define

$$F_\sigma(0, i) := \bigcup_{k=1}^{l_i} \{(f(a_{i1} \dots a_{i(k-1)}), a_{ik})\}, \quad 1 \leq i \leq d.$$

$$F_\sigma(x, i) := \{Mx + \mathbf{k} :\ \mathbf{k} \in F_\sigma(0, i)\},$$

and for $K \subseteq \mathcal{G}$,

$$F_\sigma(K) := \{F_\sigma(\mathbf{k}) :\ \mathbf{k} \in K\}.$$

Rigorously we should use $F_\sigma\{(x, i)\}$ instead of $F_\sigma(x, i)$. By the definitions of the incidence matrix and f, one has $M_{\sigma^n} = M_\sigma^n$ and $f(\sigma(U)) = M_\sigma f(U)$ for every word U, and so that $F_{\sigma^n} = F_\sigma^n$.

For a finite or infinite word $s = s_1 \dots s_n \dots$, a broken line $\overline{s}$ starting from the origin is defined as follows:

$$\overline{s} = \bigcup_{i \geq 1} \{(f(s_1 \dots s_{i-1}), s_i)\}.$$

3.2. Self-similarity of Rauzy fractals

The following lemma holds for general substitutions. (See [11] as formula (3.2).)

Lemma 3.1 ([11])**.** *For any substitution σ, it holds that*

$$Y_i = \bigcup_{j=1}^{d} \bigcup_{a_{jk}=i} (MY_j + f(a_{j1} \dots a_{j(k-1)})). \tag{3.1}$$

Set $Z_{ij} = \{f(a_{j1} \dots a_{j(k-1)}) :\ a_{jk} = i\}$, then we have that

$$Z_{ij} = \{z :\ (z, i) \in F_\sigma(O, j)\}, \tag{3.2}$$

and (3.1) becomes

$$Y_i = \bigcup_{j=1}^{d} (MY_j + Z_{ij}). \tag{3.3}$$

Recall in Theorem 1.1, $\mathcal{D}_{ij} = \{M^{-1}\pi \circ f(a_{j1} \dots a_{j(k-1)});\ a_{jk} = i\}$, we have

$$\mathcal{D}_{ij} = M^{-1}\pi(Z_{ji}) = \{M^{-1}\pi(z) :\ (z, i) \in F_\sigma(O, j)\}. \tag{3.4}$$

Proof of Theorem 1.1. (i) By the definition of F_σ, we see that every element of Y can be expressed as $y = \sum_{k=0}^{N} M^k a_k$, where a_k belongs to the finite set

$$\{f(a_{j1} \dots a_{jk}) : \ 1 \le i \le d, \ 1 \le k \le l_i\}. \tag{3.5}$$

(If s is a periodic point satisfying $\sigma^k(s) = s$, we replace σ by σ^k.) We can choose a real θ such that $|\lambda_j| < \theta < 1$ holds for all contractive eigenvalues λ_j. Since $\pi(a_k)$ are points of the contractive eigenspace P, we have

$$|\pi(y)| = \left|\sum_{k=0}^{N} M^k \pi(a_k)\right| < C \sum_{k=0}^{\infty} \theta^k = \frac{C}{1-\theta},$$

where $C > 0$ is a constant depends on the set in (3.5). So X is bounded in the ball $B(0, C/1-\theta) \subset P$ and that X_i are compact.

(ii) Taking the projection π and then taking a closure on both sides of (3.3), we obtain the set equations of the Rauzy fractals (1.1)

$$M^{-1} X_i = \bigcup_{j=1}^{d} (X_j + \mathcal{D}_{ij}) \, . \tag{3.6}$$

□

3.3. IFS and dual IFS

For a hyperbolic substitution σ, there is a natural graph IFS associated with σ.

Let V_0 be the Perron-Frobenius eigenspace of M, *i.e.*, the eigenspace corresponding to the maximal eigenvalue of M. Let $\pi'' : \mathbb{R}^d \mapsto V_0$ be the projection along the eigenspaces. Set $I_i = \{\theta \pi''(\vec{e}_i); \ 0 \le \theta \le 1\}, \ \ 1 \le i \le d$. Then it is easy to check that $\{I_i; \ 1 \le i \le d\}$ are invariant sets of the graph IFS

$$M I_i = \bigcup_{j=1}^{d} I_j + \pi'' Z_{ji}.$$

Let us build up an automaton G as following: the state set is the alphabet set $\{1, 2, \dots, d\}$, if $(z, j) \in F_\sigma(0, i)$, then we draw an edge e from state i to j and label this edge by it by z. Let us denote the label of e by d_e, *i.e.*, $d_e = z$.

In the graph G, if we associate with each edge a contraction

$$g_e(x) = M^{-1}(x + \pi''(d_e)),$$

then the invariant sets are exactly $\{I_i; \ 1 \le i \le d\}$.

Let us build up another automaton G' from G by reversing the direction of all the edges. The state set and the label of edges remain the same as G. As a corollary of Theorem 1.1, we have

Corollary 3.2. *In the graph G', if we associate with each edge a contraction*

$$g'_e(x) = A(x + \pi(d_e)),$$

then the invariant sets of the graph IFS are exactly the Rauzy fractals $X_1, \dots, X_d$.

This fact has been mentioned implicitly in [34, 3, 1, 17].

3.4. Verifying the open set condition.

Theorem 3.3.

(i) $\mathcal{D}_{ij}^n = \{\pi \circ M^{-n} z; (z,i) \in F_\sigma^n(0,j)\}$.

(ii) $\#\mathcal{D}_{ij}^n = \#\Gamma_{ij}^n$.

(iii) *There exists a constant C such that for $x \in \mathcal{D}_{ij}^n$, dist$(\pi^{-1}x, P) < C$.*

(iv) *$\#\mathcal{D}_{ij}^n$ is r-uniformly discrete.*

(v) *The open set condition holds for Rauzy fractals.*

Proof. (i) Recall that $Z_{ij} = \{z \in \mathbb{Z}^d;\ (x,i) \in F_\sigma(0,j)\}$. We set

$Z_{ij}^n = \{M^{n-1}d_{e_1} + M^{n-2}d_{e_2} + \cdots + Md_{e_{n-1}} + e_n;\ e_1 \dots e_n$ is a path from i to $j\}$.

Clearly

$$\begin{aligned}\mathcal{D}_{ij}^n &= \pi(\{M^{-n}d_{e_n} + M^{n-1}d_{e_{n-1}} + \cdots + M^{-2}d_{e_2} + M^{-1}d_{e_1};\\ &\qquad e_n \dots e_1 \text{ is a path from state } i \text{ to } j \text{ on } G'\})\\ &= \pi\left(M^{-n} Z_{ij}^n\right)\\ &= \pi \circ M^n(\{z;\ (z,i) \in F_\sigma^n(0,j)\}).\end{aligned}$$

(ii) Since the elements in $F_\sigma^n(0,j)$ form a broken line in the positive direction, all $(z,i) \in F_\sigma^n(0,j)$ are distinct, which implies that $\#\mathcal{D}_{ij}^n = \#\Gamma_{ij}^n$.

(iii) Since any $x \in \mathcal{D}_{ij}^n$ has the form

$$\pi(M^{-n}d_{e_n} + M^{n-1}d_{e_{n-1}} + \cdots + M^{-2}d_{e_2} + M^{-1}d_{e_1}),$$

we have

$$\pi^{-1}x = M^{-n}d_{e_n} + M^{n-1}d_{e_{n-1}} + \cdots + M^{-2}d_{e_2} + M^{-1}d_{e_1}.$$

Notice that all d_{e_k} are taken from a finite set (the labels of the graph G), and M^{-1} is contractive on V_1. We conclude that the distance from $\pi^{-1}x$ to P is bounded by a certain constant C, which is independent of the choices of $i, j \in \{1, \dots, d\}$ and $n \in \mathbb{N}$.

(iv) Let us denote by $\mathcal{C} := \{x :\ \text{dist}(x, P) < C\}$ the corridor near the space P. Let $B = B(O,1) \cap P$ where $B(O,1)$ is the unit ball with center O. Then there are only finite many points $z \in \mathbb{Z}^d \setminus \{0\}$ such that $\pi(z) \in B(0,1)$. Suppose $\pi(z^*)$ is the nearest point to 0 among them. Let $r = \min\{1, \pi(z^*)\}$. (If the unit ball contains only the projection of O, we set $r = 1$.) Then $\pi(\mathbb{Z}^d \cap \mathcal{C})$ is r-uniformly discrete, and hence its subsets $\mathcal{D}_{ij}^n$ are also r-uniformly discrete for all $1 \le i, j \le d$, $n \ge 1$.

(v) Therefore, the open set condition holds by Theorem 2.1. This proves Theorem 1.2. □

4. Domain-exchange transformation on X

4.1. Strong coincidence condition

Let σ be a substitution over $\mathcal{A} = \{1, \dots, d\}$. Two distinct letters $i, j \in \mathcal{A}$ are said to have *strong coincidence* if there exist integers k, n such that $\sigma^n(i)$ and $\sigma^n(j)$ have the same kth letter, and the prefixes of length $k-1$ of $\sigma^n(i)$ and $\sigma^n(j)$

have the same image under the abelianization map f. A substitution σ is said to satisfy the *strong coincidence condition* if i, j have coincidence for any $i, j \in \mathcal{A}$. (See [16, 1, 6, 17].)

It is conjectured that all Pisot substitutions satisfy the strong coincidence condition ([6, 17]) and the conjecture is confirmed for Pisot substitutions over two letters [6]. However, a hyperbolic substitution may not satisfy the strong coincidence condition. Let τ be the substitution

$$\tau : 1 \to 12,\ 2 \to 3,\ 3 \to 24,\ 4 \to 1.$$

We conjecture that the pair $2, 3$ do not have strong coincidence.

A measure μ on X is called translation invariant (or a Haar type measure) if $\mu(B_1) = \mu(B_2)$ holds provided B_1 and B_2 are Borel subset of X and $B_1 = B_2 + v$ for some vector v.

Let $\alpha = \dim_w X$ be the Hausdorff dimension of X w.r.t. the pseudo norm $w = w_M$. Let

$$\mu = a\mathcal{H}_w^\alpha|_X$$

where $a = 1/\mathcal{H}_w^\alpha(X)$. Clearly μ is translation invariant on X.

Proposition 4.1. *Let σ be a hyperbolic unimodular substitution with partial atomic surfaces $X_1, \ldots, X_d$. If σ satisfies the strong coincidence condition, then $\mu(X_i \cap X_j) = 0$ for $i \neq j$.*

Proof. Since the letters i and j have strong coincidences, there is a letter $k \in \mathcal{A}$, an integer $n > 0$ and a lattice point $z \in \mathbb{Z}^d$ such that $(z, k) \in F_\sigma^n(O, i) \cap F_\sigma^n(O, j)$. Hence $(z, k) \in F_{\sigma^n}(O, i) \cap F_{\sigma^n}(O, i)$. By the definition of $\mathcal{D}_{ki}$ we have that $\pi(z) \in \mathcal{D}_{ki}^n \cap \mathcal{D}_{kj}^n$. This implies that both $X_i + \pi(z)$ and $X_j + \pi(z)$ are pieces of the subdivision of $M_\sigma^n(X_k)$.

Hence, by the OSC, we have that $\mathcal{H}_w^\alpha\left((X_i + \pi(z)) \cap (X_j + \pi(z))\right) = 0$, and so that $\mu(X_i \cap X_j) = 0$. □

4.2. Domain-exchange transformation

Actually, there is another way to tile X by X_i. Consider

$$Y_i' = \{f(s_1 s_2 \ldots s_n) : \ s_n = i\},$$

then Y_i' is another partition of the stair determined by the fixed point. Clearly $Y_i' = Y_i + e_i$. Let $X_i' = \overline{Y_i'}$, then

$$X_i' = X_i + \pi(e_i) \text{ and } X = \cup_{i=1}^d X_i',$$

and the union is non-overlapping in measure μ since $\mu(X_i) = \mu(X_i')$ and the total measure is unchanged.

Accordingly, a *domain-exchange transformation* $D : X \mapsto X$ can be defined by

$$D(x) = x + \pi(e_k), \qquad \text{if } x \in X_k \text{ and } x \notin \bigcup_{j=1}^{k-1} X_j.$$

For Pisot substitutions, this transformation was defined and studied in [1].

Theorem 4.2. (X, D, μ) *is a measure preserving dynamical system.*

Proof. Actually, we shall prove that for any Borel set B, $\mu(D(B)) = \mu(B)$, which is stronger than that $\mu(D^{-1}(B)) = \mu(B)$.

Let $U = \bigcup_{i,j}(X_i \cap X_j)$ and set $\tilde{X}_k := X_k \setminus U = X_k \setminus (\cup_{j \neq k} X_j)$. Clearly $\mu(X) = \sum_{k=1}^{d} \mu(\tilde{X}_k)$. By the translation invariant of D, we have

$$\mu(D(\tilde{X}_i)) = \mu(\tilde{X}_i + \pi(e_i)) = \mu(\tilde{X}_i).$$

Since $D(\tilde{X}_i) \subset X'_i$ and $X = \cup_{i=1}^{d} X'_i$ is a disjoint union in measure μ, we have that $D(\tilde{X}_i)$ are also disjoint in measure μ.

Set $B_k := B \cap \tilde{X}_k$. Then $\mu(B) = \sum_{k=1}^{d} \mu(B_k)$ and $D(B_k) = B_k + \pi(e_k) \subset D(\tilde{X}_k)$. Hence $D(B_k) \cap D(B_{k'})$ is a subset of $D(\tilde{X}_k) \cap D(\tilde{X}_{k'})$, and so that $\mu(D(B_k) \cap D(B_{k'})) = 0$. Therefore,

$$\mu(D(B)) = \bigcup_{k=1}^{d} \mu(D(B_k)) = \bigcup_{k=1}^{d} \mu(B_k) = \mu(B). \qquad \square$$

5. Realization of the substitution dynamical system

5.1. Isomorphism of measure-preserving transformations

The following definition can be found in [35] p. 57.

Definition 5.1. Suppose $(X_1, \mathcal{B}_1, m_1)$ and $(X_2, \mathcal{B}_2, m_2)$ are probability spaces together with measure-preserving transformations

$$T_1 : X_1 \mapsto X_1,\ T_2 : X_2 \mapsto X_2.$$

We say that T_1 is *isomorphic* to T_2 if there exist $M_1 \in \mathcal{B}_1$, $M_2 \in \mathcal{B}_2$ with $m_1(M_1) = 1$, $m_2(M_2) = 1$ such that

(i) $T_1 M_1 \subset M_1$, $T_2 M_2 \subset M_2$, and

(ii) there is an invertible measure-preserving transformation

$$\phi : M_1 \mapsto M_2 \text{ with } \phi T_1(x) = T_2 \phi(x)\ \forall x \in M_1.$$

In (ii) the set M_i $(i = 1, 2)$ is assumed to be equipped with the σ-algebra $M_i \cap \mathcal{B}_i = \{M_i \cap B;\ B \in \mathcal{B}_i\}$ and the restriction of the measure m_i to this σ-algebra.

Clearly, if two systems are isometric, then they are spectrally isomorphic. (See [35] Ch. 2.)

5.2. Construction of conjugacy

Recall that (Ω, T, ν) is the substitution dynamical system, where $\Omega = \overline{\{T^n(s); n \geq 0\}}$ is the orbit closure of the fixed point s of σ, T is the shift operator and ν is the unique ergodic probability measure.

We define a projection $\rho :\ \Omega \mapsto X$. First, set

$$\rho(T^n(s)) = \rho(s_n s_{n+1} \dots) = \pi \circ f(s_0 s_1 \dots s_{n-1}),$$

where s_0 denotes the empty word and s_1 is the initial letter of the fixed point s.

Lemma 5.2. *ρ is uniformly continuous on the set $\{T^n(s);\ n \geq 0\}$.*

Proof. Suppose $T^n(s)$ and $T^m(s)$ are very close to each other, that is,

$$s_n s_{n+1} \dots s_{n+L} = s_m s_{m+1} \dots s_{m+L} \tag{5.1}$$

for very large L.

Primitive aperiodic substitutions are known to be bilaterally recognizable [26]. Hence we can write (5.1) in the form

$$s_n s_{n+1} \dots s_{n+L} = s_m s_{m+1} \dots s_{m+L} = W_1 \sigma^p(i) W_2, \tag{5.2}$$

where $i \in \mathcal{A}$, $W_1, W_2 \in \mathcal{A}^*$ and

$$s_0 s_1 \dots s_{m-1} W_1 = \sigma^p(U_1), \quad s_0 s_1 \dots s_{n-1} W_1 = \sigma^p(U_2)$$

for some words $U_1, U_2 \in \mathcal{A}^*$. Moreover, $p = p(L)$ depends only on σ and L, and p tends to $+\infty$ when $L \to +\infty$.

Set $k = |W_1|$, we have

$$\begin{aligned}
|\rho(T^n s) - \rho(T^m s)| &= |\rho(T^{n+k} s) - \rho(T^{m+k} s)| \\
&= |\pi \circ f(s_0 s_1 \dots s_{n+k}) - \pi \circ f(s_0 s_1 \dots s_{m+k})| \\
&= |\pi \circ f(\sigma^p(U_1)) - \pi \circ f(\sigma^p(U_2))| \\
&= |\pi \circ M^p(f(U_1)) - \pi \circ M^p(f(U_2))| \\
&\leq |M^p \circ (\pi(f(U_1)) - \pi(f(U_2))\,| \\
&\leq 2\text{diam }(M^p X) \leq 2\theta^p \text{diam } X,
\end{aligned}$$

where $|\lambda_j| < \theta < 1$ holds for every contractive λ_j. Therefore, when L is large, p must be large and so that $2\theta^p$diam X is small. This proves the uniform continuity of ρ. □

Since $T^n(s)$ is dense in Ω, hence ρ can be extended to Ω by continuity, and ρ is still continuous on Ω. Therefore

$$\rho:\ \Omega \to X$$

is well defined and it is an onto map by the definition of Rauzy fractal X.

It is easy to show that the following diagram is commutative.

$$\begin{tikzcd}
\Omega \arrow[r, "T"] \arrow[d, "\rho"'] & \Omega \arrow[d, "\rho"] \\
X \arrow[r, "D"'] & X
\end{tikzcd}$$

Proposition 5.3. *The following holds.*

(i) $\rho[i] = X_i$.

(ii) $\rho \circ T = D \circ \rho$.

(iii) *For $[i_1 \dots i_n] \subset \Omega$, we have that $\rho[i_1 \dots i_n] = X_{i_1} \cap D^{-1}(X_{i_2}) \cap \cdots \cdots \cap D^{-n+1}(X_{i_n})$.*

(iv) *$\rho^{-1} : \tilde{X} \mapsto \tilde{\Omega}$ is a measurable mapping.*

Theorem 1.3. *If σ satisfies the strong coincidence condition, then (X, D, μ) is isomorphic to the substitution dynamical system (Ω, T, ν).*

Proof. The restriction of measure μ on $\tilde{X}$ is still a probability measure and we still denote it by μ. Since ρ^{-1} is measurable, $\nu' = \mu \circ \rho$ is the induced Borel probability measure on $\tilde{\Omega}$. Let us extend ν' to Ω by setting $\nu'(\Omega \setminus \tilde{\Omega}) = 0$. We obtain a Borel probability measure on Ω and we still denote it by ν'. It is easy to show that after extension, we still have $\nu' = \mu \circ \rho$.

Now we show that ν' is T-invariant.

$$\begin{aligned} \nu'(T(H)) &= \mu \circ \rho \circ T(H) = \mu \circ D \circ \rho(H) \\ &= \mu \circ \rho(H) = \nu'(H). \end{aligned}$$

Since (Ω, T) is uniquely ergodic, we obtain that $\nu' = \nu$. □

References

[1] P. Arnoux and S. Ito, *Pisot substitutions and Rauzy fractals.* Bull. Belg. Math. Soc. **8** (2001), no. 2, 181–207.

[2] S. Akiyama, *Self affine tiling and Pisot numeration system.* In Number theory and its applications, ed. by Győry and S. Kanemitsu, Kluwer (1999), 7–17.

[3] S. Akiyama, *On the boundary of self affine tilings generated by Pisot numbers.* J. Math. Soc. Japan **54** (2002), no. 2, 283–308.

[4] C. Bandt, *Self-similar sets* 5. *Integer matrices and fractal tilings of* $\mathbb{R}^n$. Proc. AMS **112** (1991), no. 2, 549–562.

[5] C. Bandt and S. Graf, *Self-similar sets* 7. *A characterization of self-similar fractals with positive Hausdorff measure.* Proc. AMS **114** (1992), 995–1001.

[6] M. Barge and B. Diamond, *Coincidence for substitutions of Pisot type.* Bull. Soc. Math. France **130** (2002), no. 4, 619–626.

[7] M. Baake and U. Grimm, *A guide to quasicrystal literature.* In Directions in mathematical quasicrystals, CRM Monogr. Ser.,(2000), 371–373.

[8] V. Canterini and A. Siegel, *Geometric representation of primitive substitutions of Pisot type.* Trans. AMS **353** (2001), no. 12, 5121–5144.

[9] F.M. Dekking, *The spectrum of dynamical systems arising from substitutions of constant lenght.* Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **41** (1977/78), no. 3, 221–239.

[10] H. Ei and S. Ito, *Tilings from some non-irreducible, Pisot substitutions.* Discrete Math. Theor. Comput. Sci. **7** (2005), no. 1, 81–121 (electronic).

[11] H. Ei, S. Ito and H. Rao, *Atomic surfaces, tilings coincidences II: Reducible case.* Ann. Inst. Fourier (Grenoble) **56** (2006), no. 7, 2285–2313.

[12] K. Falconer, *Techniques in fractal geometry.* John Wiley & Sons Ltd., Chichester, 1997

[13] P. Fogg, *Substitutions in dynamics, arithmetics and combinatorics.* Lecture Notes in Math., 1794, Springer, Berlin, 2002.

[14] X.G. He and K.S. Lau, *On a generalized dimension of self-affine fractals.* Math. Nachr. **281** (2008), no. 8, 1142–1158.

[15] C. Holton and L. Zamboni, *Geometric realizations of substitutions.* Bull. Soc. Math. France **126** (1998), no. 2, 149–179.

[16] B. Host, Special substitutions on 2 letters. Unpublished manuscript.

[17] S. Ito and H. Rao, *Atomic surfaces, tilings and coincidences I. Irreducible case.* Israel J. Math. **153** (2006), 129–156.

[18] R. Kenyon, *Self-replicating tilings.* in Symbolic dynamics and its applications, Contemporary mathematics series. Vol. 135 (P. Walters, ed.), American Mathematical Society, Providence, RI, 1992.

[19] P.G. Lemarié-Rieusset, *Projecteurs invariants, matrices de dilatation, ondelettes et analyses multi-résolutions.* Rev. Math. Iberoamer. **10** (1994), 283–348.

[20] J. Lagarias J. and Y. Wang, *Self-affine tiles in* $\mathbb{R}^n$*.* Adv. Math. **121** (1996), 21–49.

[21] J. Lagarias J. and Y. Wang, *Substitution Delone sets.* Discrete Comput. Geom. **29** (2003), no. 2, 175–209.

[22] J. Luo and Y.M. Yang, *On single-matrix graph directed iterated function systems.* Preprint 2008.

[23] J.M. Luck, C. Godreche, A. Janner, and T. Janssen, *The nature of the atomic surfaces of quasiperiodic self-similar structures.* J. Phys. A: Math. Gen. **26** (1993), 1951–1999.

[24] C. McMullen, *The Hausdorff dimension of general Sierpinski carpets.* Nagoya Math. J. **96** (1984), 1–9.

[25] P. Michel, *Stricte ergodicité d'ensembles minimaux de substitutions.* C.R. Acad. Sc. Paris **278** (1974), 811–813.

[26] B. Mossé. *Puissances de mots et reconnaisabilité des points fixes d'une substitution.* Theor. Comp. Sci. **99** (1992), no. 2, 327–334.

[27] R.D. Mauldin and S. Williams, *Hausdorff dimension in graph directed constructions.* Trans. Amer. Math. Soc. **309** (1988), no. 2, 811–829.

[28] B. Praggastis, *Numeration systems and Markov partitions from self-similar tilings.* Trans. Amer. Math. Soc. **351** (1999), no. 8, 3315–3349.

[29] M. Queffelec, *Substitution Dynamical Systems – Spectral Analysis.* Lecture Notes in Math. **1294** (1987), Springer-Verlag, New York.

[30] G. Rauzy, *Nombres algébriques et substitutions.* Bull. Soc. Math. France **110** (1982), no. 2, 147–178.

[31] A. Schief, *Separation properties for self-similar sets.* Proc. Amer. Math. Soc. **122** (1994), no. 1, 111–115.

[32] A. Siegel, *Pure discrete spectrum dynamical system and periodic tiling associated with a substitution.* Ann. Inst. Fourier (Grenoble) **54** (2004), no. 2, 288–299.

[33] V. Sirvent and Y. Wang, *Self-affine tiling via substitution dynamical systems and Rauzy fractals.* Pacific J. Math. **206** (2002), no. 2, 465–485.

[34] W.P. Thurston, *Groups, tilings, and finite state automata.* (Boulder, CO, 1989) Amer. Math. Soc. Colloq. Lectures.

[35] P. Walters, *An introduction to ergodic theory.* Springer-Verlag, New York Inc. 1982.

Maki Furukado
Faculty of Business Administration
of Yokohama University
79-4, Tokiwadai, Hodogaya-ku
Yokohama, 240-8501, Japan
e-mail: furukado@ynu.ac.jp

Shunji Ito
Graduate School of Natural Science
& Technology of Kanazawa University
Kakuma-machi
Kanazawa, 920-1192, Japan
e-mail: ito@t.kanazawa-u.ac.jp

Hui Rao
Mathematical Department
of Tsinghua University
Beijing 100084, China
e-mail: hrao@math.tsinghua.edu.cn

Progress in Probability, Vol. 61, 269–284

Random Cantor Sets and Their Projections

Michel Dekking

Abstract. We discuss two types of random Cantor sets, M-adic random Cantor sets, and Larsson's random Cantor sets. We will discuss the properties of their ninety and fortyfive degree projections, and for both we give answers to the question whether the algebraic difference of two independent copies of such sets will contain an interval or not.

Mathematics Subject Classification (2000). Primary 28A80 Secondary 60J80, 60J85.

Keywords. Random fractals, difference of Cantor sets, Palis-Takens conjecture, multitype branching processes.

1. Introduction

We start with an example of the kind of random Cantor sets that we will discuss. The classical triadic Cantor set is recursively obtained by dividing the unit interval in the three intervals $[0,1/3],[1/3,2/3]$ and $[2/3,1]$, and then discarding the middle interval. The two remaining intervals form C^1, and are subdivided into three intervals of length $1/9$, and in each of them the middle interval is discarded. Continuing in this way one obtains a set C^n of 2^n intervals of length 3^{-n} in the n^{th} step, and the classical triadic Cantor set is the intersection of the C^n. In the random version, one flips independently three coins for each of the intervals $[0,1/3],[1/3,2/3]$ and $[2/3,1]$. If a coin turns up heads then the corresponding interval is discarded, if tails turns up then the corresponding interval is a member of C^1. Next, each interval of C^1 is subdivided into three intervals of length $1/9$, and the same coin flipping procedure is applied to these three intervals. In this way one obtains a decreasing sequence of random sets C^n (C^n might be empty), and the triadic random Cantor set C is the intersection of the C^n.

Algebraic differences (or sums) of deterministic Cantor sets have received a lot of attention lately, see, e.g., [1], [2], [25], [12]. Here the *algebraic difference* (or: *arithmetic difference*) of two arbitrary sets $A, B \subseteq \mathbb{R}$ is defined as the set of those real numbers that can be formed as the difference of a number from the first set and a number from the second set:

$$A - B := \{x - y : x \in A, y \in B\}.$$

In this paper we will focus on the question whether the difference of two *random* Cantor sets C_1 and C_2 contains an interval or not. The condition $\dim_H C_1 + \dim_H C_2 \geq 1$, is a *necessary* condition for the algebraic difference $C_1 - C_2$ to contain an interval. A conjecture by Palis and Takens ([23]) states that if

$$\dim_H C_1 + \dim_H C_2 > 1, \tag{1}$$

then *generically* it should be true that $C_1 - C_2$ contains an interval. Excluding the case $\dim_H C_1 + \dim_H C_2 = 1$ the Palis-Takens conjecture thus essentially states that the necessary condition regarding the sum of Hausdorff dimensions would also be a *sufficient* condition.

For generic dynamically generated *non-linear* Cantor sets this was proved in 2001 by de Moreira and Yoccoz ([4]). The problem is open for generic linear Cantor sets.

The problem was put into a probabilistic context by Per Larsson in his thesis [20] (see also [19]). He considers a two parameter family of random Cantor sets $C_{a,b}$, and claims to prove that the Palis conjecture holds for all relevant choices of the parameters a and b (see Section 5).

We will consider the problem for M-adic random Cantor sets (where the Palis conjecture may fail to hold in general). A simple observation (see Section 4) gives that studying the difference is the same as studying a 45° projection. We will therefore first (in Section 3) give similar results for the easier case of orthogonal projections.

2. Random Cantor sets

We will introduce random M-adic Cantor sets, defined for any integer $M \geq 2$, the *base* of the Cantor set. The randomness resides in the process of deletion of intervals. See Section 5 for another type of random Cantor sets, where the *position* of the intervals is random.

2.1. Trees

The construction of M-adic Cantor sets is intimately related to the notion of M-ary trees.

Let $M \geq 2$ be an integer. An M-ary tree is a tree in which every node has precisely M children. See Figure 1 for an example. The nodes are conveniently identified with strings over an alphabet of size M; we use the alphabet $\mathbb{A} := \{0, \dots, M-1\}$.

Strings over $\mathbb{A}$ of length n are denoted as $i_1 \dots i_n$, where $i_1, \dots, i_n \in \mathbb{A}$. The empty string is denoted by $\emptyset$ and has length 0.

The *M-ary tree* $\mathcal{T}$ is defined as the set of all strings over the alphabet $\mathbb{A}$. The root node is the empty string $\emptyset$. The *children* of each node $i_1 \dots i_n \in \mathcal{T}$ are the nodes $i_1 \dots i_n i_{n+1}$ for all $i_{n+1} \in \mathbb{A}$. The *level* of a node corresponds to its length as a string. For each $n \geq 0$, the n^{th} *level* of the tree consists of all nodes at level n, and it is denoted by $\mathcal{T}_n$. It thus holds that

$$\mathcal{T} = \bigcup_{n\geq 0} \mathcal{T}_n = \bigcup_{n\geq 0} \bigcup_{i_1\in\mathbb{A}} \cdots \bigcup_{i_n\in\mathbb{A}} \{i_1 \dots i_n\}.$$

2.2. Random M-adic Cantor sets

Let $\mathbf{p} := (p_0, \dots, p_{M-1}) \in [0,1]^M$, be a vector of probabilities (in general *not* a probability vector). We introduce a probability measure $\mathbb{P}_{\mathbf{p}}$ on the space of labeled trees, by giving each node $i_1 \dots i_n$ a random label $X_{i_1\dots i_n}$ which will be 0 or 1. The probability measure is defined by requiring that the $X_{i_1\dots i_n}$ are independent Bernoulli random variables, with $\mathbb{P}_{\mathbf{p}}(X_\emptyset = 1) = 1$, and for $n \geq 1$ and $i_1 \dots i_n \in \mathbb{A}^n$

$$\mathbb{P}_{\mathbf{p}}(X_{i_1\dots i_n} = 1) = p_{i_n}.$$

The randomly labeled tree determines a random Cantor set in [0,1] in the following way. Define for all $i_1 \dots i_n \in \mathcal{T}$

$$I_{i_1\dots i_n} := \left[\frac{i_1}{M} + \frac{i_2}{M^2} + \cdots + \frac{i_n}{M^n}, \frac{i_1}{M} + \frac{i_2}{M^2} + \cdots + \frac{i_n}{M^n} + \frac{i_n + 1}{M^n}\right].$$

The n^{th} level approximation C^n of the random Cantor set C is a union of such n^{th} level M-adic intervals selected by the *survival* sets $S_n \subset \mathcal{T}_n$ defined by

$$S_n = \{i_1 \dots i_n : X_{i_1} = X_{i_1 i_2} = \cdots = X_{i_1\dots i_n} = 1\}.$$

The random Cantor set C is

$$C = \bigcap_{n=1}^{\infty} C^n = \bigcap_{n=1}^{\infty} \bigcup_{i_1\dots i_n\in S_n} I_{i_1\dots i_n}.$$

The classical triadic Cantor set can be obtained in this way by taking $M = 3$, and the vector

$$\mathbf{p} := (p_0, p_1, p_2) = (1, 0, 1).$$

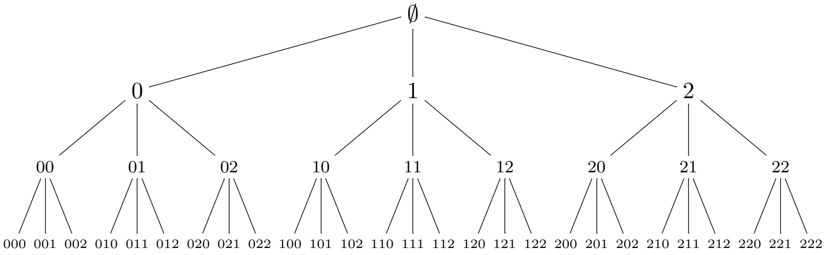

FIGURE 1. A graphical representation of the first four levels – $\mathcal{T}_0$, $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ – of the 3-ary (ternary) tree $\mathcal{T}$.

A well-known (see, e.g., [3]) example is *M-adic fractal percolation with parameter p*, which is obtained by choosing a number p in $[0,1]$, and taking the vector $\mathbf{p}$ of length M equal to

$$\mathbf{p} = (p, p, \dots, p)\,.$$

This set is called also *Mandelbrot percolation* or *canonical curdling* (by Mandelbrot himself). The term percolation actually refers to a two-dimensional version, which we will discuss next (see also Figure 2).

2.3. Higher dimensions

Obviously the idea of M-adic Cantor sets can be extended to d-dimensional Euclidean space, by considering the M^d-adic tree, whose nodes correspond to d-dimensional level n M-adic hypercubes

$$I^{(n)}_{\underline{j}_1,\dots,\underline{j}_d} = I^{(n)}_{\underline{j}_1} \times \cdots \times I^{(n)}_{\underline{j}_d}, \quad \underline{j}_k \in \mathbb{A}^n.$$

The limit set C in $[0,1]^d$ is then defined by an intersection of approximating sets C^n as above. In accordance with the structure of hypercubes we arrange the level n symbols in matrices of size $M^n \times M^n \times \cdots \times M^n$.

In the sequel we shall for simplicity of notation mainly consider the case $d = 2$. As a (deterministic) example, let $M = 3$ and define $\mathbf{p}^S$ by

$$\mathbf{p}^S = \begin{pmatrix} p^S_{00} & p^S_{01} & p^S_{02} \\ p^S_{10} & p^S_{11} & p^S_{12} \\ p^S_{20} & p^S_{21} & p^S_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

The limit set $C(\mathbf{p}^S)$ equals the *Sierpinski carpet.* This set is not a Cantor set in the usual sense, since it is not totally disconnected. However, in the sequel we will continue to use the name Cantor set, were we do not refer to its topological properties, but rather to the similarity of its construction with that of the classical triadic Cantor set. It is in fact an open problem to determine for two-dimensional fractal percolation with parameter p for which p the set is (almost surely) totally disconnected, and for which p it is not ([3]).

Triadic fractal percolation in the plane is determined by a p with $0 \le p \le 1$, and the matrix

$$\mathbf{p} = \begin{pmatrix} p & p & p \\ p & p & p \\ p & p & p \end{pmatrix}.$$

See Figure 2 for realisations of the first six approximants to C, where $M = 2$.

As a final example take $d = 2$, $M = 3$ and a p in $[0,1]$, and choose $\mathbf{p}^{S,p}$ according to

$$\mathbf{p}^{S,p} = \begin{pmatrix} p & p & p \\ p & 0 & p \\ p & p & p \end{pmatrix}.$$

In this way we obtain what we call the *random Sierpinski carpet* with parameter p. See Figure 3 for a realisation of C^5 of this carpet.

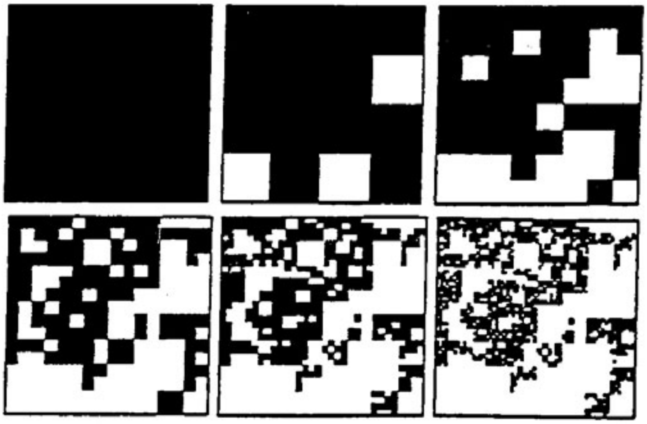

FIGURE 2. Realizations $C^1, \ldots, C^6$ of base 2 fractal percolation.

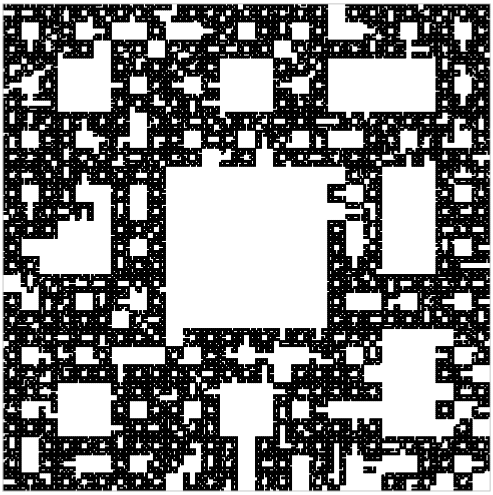

FIGURE 3. A realisation of C^5 for the random Sierpinski carpet ($p = .9$).

2.4. Joint survival distributions

So far we have considered M-adic random Cantor sets where the intervals are picked or discarded independently of each other. However, if we collapse two steps of the construction into one, passing to a M^2-adic Cantor set generating the sequence $C^{2n}, n = 1, 2, \ldots$, which (in distribution) yields the same set

$$C = \bigcap_{n=0}^{\infty} C^{2n},$$

then this independence is lost. For example, in fractal percolation in base $M = 3$,

$$\mathbb{P}_{\mathbf{p}}(I_{00} \subset C^2)\mathbb{P}_{\mathbf{p}}(I_{01} \subset C^2) = p^4 \neq \mathbb{P}_{\mathbf{p}}(I_{00} \subset C^2,\, I_{01} \subset C^2) = p^3.$$

Since this *order 2* Cantor set – and higher orders – will be important later, it is necessary to generalize the definition of random Cantor sets. Let μ be a probability measure on the collection of subsets of $\mathbb{A}$. We call μ the *joint survival distribution.* This measure induces a probability measure $\mathbb{P}_\mu$ on the space of labeled trees, where we label each node $i_1 \dots i_n \in \mathcal{T}$ with $X_{i_1\dots i_n} \in \{0,1\}$, by requiring that $\mathbb{P}_\mu\left(X_\emptyset = 1\right) = 1$ and that the sets

$$\left\{i_{n+1} \in \mathbb{A} : X_{i_1\dots i_n i_{n+1}} = 1\right\} \sim \mu \tag{2}$$

are independent (and identically) distributed for all $i_1 \dots i_n \in \mathcal{T}$.

For a joint survival distribution μ the vector of *marginal probabilities* $\mathbf{p} := (p_0, \dots, p_{M-1})$ is defined by

$$p_i := \mathbb{P}_\mu\left(X_i = 1\right), \tag{3}$$

for all $i \in \mathbb{A}$. Note that these marginal probabilities do not need to sum up to 1. It is remarkable that many properties of C are already determined by these marginal probabilities, rather than by the complete joint survival distribution. We give two simple examples of this phenomenon in the next section.

2.5. Two basic structural properties

First we treat the question under which conditions C is empty. The number of level n intervals selected in C^n, $\mathrm{Card}(S_n)$, is a branching process with as offspring distribution the distribution of $\mathrm{Card}(S_1)$. Since the C^n are non-increasing, $C = \emptyset$ if and only if the branching process $(\mathrm{Card}(S_n))$ dies out. Note that $\mathbb{E}_\mu[\mathrm{Card}(S_1)] = \|\mathbf{p}\|_1$.

It follows from well-known results in branching process theory that if

$$\|\mathbf{p}\|_1 \leq 1 \text{ and } \mathbb{P}_\mu\left(\mathrm{Card}(S_1) = 1\right) \neq 1,$$

then $C = \emptyset$ almost surely, and if on the other hand

$$\|\mathbf{p}\|_1 > 1 \text{ or } \mathbb{P}_\mu\left(\mathrm{Card}(S_1) = 1\right) = 1,$$

then $C \neq \emptyset$ with positive probability.

The next question is: what is the dimension of C? It is well known (see, e.g., [18], [13]) that the Hausdorff dimension $\dim_H$ and the box dimension $\dim_B$ of the random Cantor set C are given by

$$\dim_H C = \dim_B C = \frac{\log \mathrm{Card}(S_1)}{\log M} = \frac{\log(\sum_{i\in\mathbb{A}} p_i)}{\log M} = \frac{\log \|\mathbf{p}\|_1}{\log M},$$

almost surely on $\{C \neq \emptyset\}$.

3. Orthogonal projections of M-adic random Cantor sets

Random Cantor sets are already considered in Mandelbrot's book [21]. There he uses 5-adic fractal percolation (called random curdling) in dimension $d = 3$ to implement Hoyle's model for galaxies. He gives the reader an idea of this galaxy model by presenting plane projections, and remarks that "it is not rare that two contributing cubes project on the same square", but "that in the limit the projections of two points almost never coincide". This is poetically phrased as "The dust is so sparse as to leave space essentially transparent".

This raises several questions about the size and the structure of an orthogonal projection of an M-adic random Cantor set, as for instance "When does the dimension not drop?", "When is the Lebesgue measure of the projection positive?" (see, e.g., [6]). Here we will focus on the question

"When will the projection contain a ball, i.e., have non-empty interior?"

For orthogonal projections of M-adic random Cantor sets this question has been answered by Falconer and Grimmett. We consider projections in $\mathbb{R}^d$ on the first e coordinates, where $1 \le e < d$. For $\vec{j} = (j_1, \dots, j_e) \in \mathbb{A}^e$ let

$$N(\vec{j}) = \mathrm{Card}\{\vec{i} = (i_1, \dots, i_d) \in S_1 : (i_1, \dots, i_e) = \vec{j}\}, \quad \text{and} \quad m_{\vec{j}} = \mathbb{E}_\mu[N(\vec{j})].$$

These are the number of surviving hypercubes in C^1 which project onto the subcube of $\mathbb{R}^e$ corresponding to the vector $\vec{j} = (j_1, \dots, j_e)$, and their expectation. We also need

$$N_{\min} := \min_{\vec{j} \in \mathbb{A}^e} N(\vec{j}).$$

Theorem 1. ([14]+[15]) *Let C be an M-adic random Cantor set in $\mathbb{R}^d$ generated by a joint survival distribution μ, and let π_e be the orthogonal projection on $\mathbb{R}^e$, where $1 \le e < d$.*

(a) *If for all $\vec{j} \in \mathbb{A}^e$ $m_{\vec{j}} > 1$, then $\pi_e C$ contains a ball almost surely on $\{C \neq \emptyset\}$ if and only if $\mathbb{P}_\mu (N_{\min} \ge 1) = 1$ or $\mathbb{P}_\mu (N_{\min} \ge 2) > 0$.*

(b) *If there exists a $\vec{j} \in \mathbb{A}^e$ with $m_{\vec{j}} < 1$, then $\pi_e C$ contains no ball almost surely.*

As an example, let $d{=}2$, $e{=}1$, $M = 3$, and let μ give probability 1/3 to the three sets consisting of two full columns (cf. Figure 4).

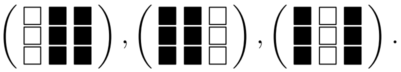

FIGURE 4. Sets $\mathbb{A}^2 \setminus \{(i, 0), (i, 1), (i, 2)\}$, for $i = 0, 1, 2$ obtaining probability 1/3.

In this example $m_0 = m_1 = m_2 = 2$, but $N_{\min} \equiv 0$. Thus part (a) of Theorem 1 implies that $\pi_1 C$ does not contain an interval (i.e., a one-dimensional ball). It is interesting to remark that $\pi_1 C$ *does* have positive Lebesgue measure, since $m_0 m_1 m_2 = 8$, which is larger than 1 (see Theorem 8 in [5]).

4. Fortyfive degree projections and differences

The algebraic difference $A - B$ of two sets A and B can be seen (modulo a dilation by $\sqrt{2}$) as a projection under 45° of the Cartesian product $A \times B$. See Figure 5 for an illustration of this.

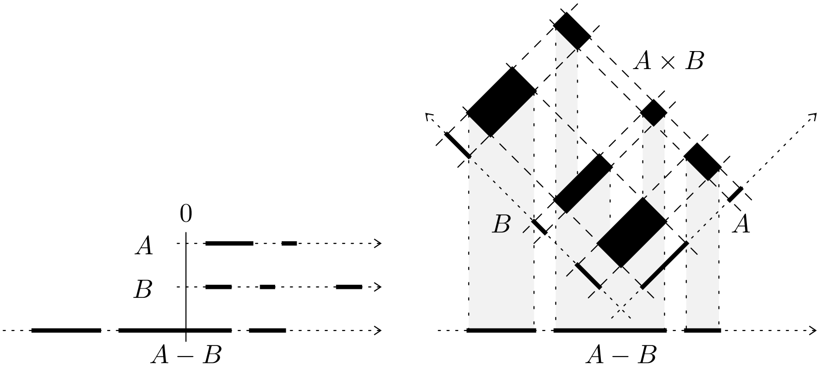

FIGURE 5. The algebraic difference $A - B$ of $A, B \subseteq \mathbb{R}$ can be interpreted as a projection of the product set $A \times B$ by the map $(x, y) \mapsto x - y$.

We consider the algebraic difference $C_1 - C_2$ between two independent random M-adic Cantor sets C_1 and C_2. In general, we denote the joint survival distribution of C_1 by μ and that of C_2 by λ. The corresponding marginal distributions will be denoted by $\mathbf{p}$ and $\mathbf{q}$ respectively. Sometimes, we will restrict ourselves to the (*marginal*) *symmetric case*, where $\mathbf{p} = \mathbf{q}$, but in general we will allow μ and λ to have different marginal probabilities.

The algebraic difference $C_1 - C_2$ is defined on the product space of the probability spaces of C_1 and C_2. We will use $\mathbb{P} := \mathbb{P}_\mu \times \mathbb{P}_\lambda$ to denote the corresponding product measure and $\mathbb{E}[\cdot]$ to denote expectations with respect to this probability.

4.1. Correlation coefficients

Define the *cyclic cross-correlation coefficients*

$$\gamma_k := \sum_{i=0}^{M-1} p_i q_{i-k}, \tag{4}$$

where the indices of p should be taken modulo M, and $k \in \mathbb{A}$. This definition is extended to $k \in \mathbb{Z}$ by setting $\gamma_{k+iM} := \gamma_k$ for all $i \in \mathbb{Z} \setminus \{0\}$. For the symmetric case $\mathbf{p} = \mathbf{q}$, these coefficients are called the *cyclic auto-correlation coefficients*. For brevity, however, we will use the shorter term *correlation coefficients*, without making a distinction between the symmetric and the non-symmetric case.

4.2. A partial result

The main theorem in [7] links the γ_k – the correlation coefficients defined in Section 4.1 – to the interval-or-not question. In the setting of general joint survival distributions μ and λ, the condition stated below is sufficient for the theorem to hold.

For a joint survival distribution $\mu : 2^{\mathbb{A}} \to [0,1]$ we define its *marginal support*:

$$\mathrm{Supp_m}(\mu) := \bigcup \{S \subseteq \mathbb{A} : \mu(S) > 0\}\,. \tag{5}$$

In other words, the marginal support is the set of $i \in \mathbb{A}$ for which it holds that $p_i = \mathbb{P}_\mu(X_i = 1) > 0$. For example take $M = 4$ and μ defined by $\mu(\{0,3\}) = \mu(\{1,3\}) = \mu(\{3\}) = \frac{1}{3}$, then the marginal support is $\mathrm{Supp_m}(\mu) = \{0,1,3\}$.

We say a joint survival distribution $\mu : 2^{\mathbb{A}} \to [0,1]$ satisfies the *joint survival condition* if it assigns a positive probability to its marginal support:

$$\mu(\mathrm{Supp_m}(\mu)) > 0. \tag{6}$$

Joint survival distributions that correspond to M independent Bernoulli variables satisfy the joint survival condition since we have

$$\mu(\mathrm{Supp_m}(\mu)) = \prod_{i \in \mathrm{Supp_m}(\mu)} p_i > 0$$

for them. In the example above, where $\mathrm{Supp_m}(\mu) = \{0,1,3\}$, the joint survival condition is not satisfied, because the set $\{0,1,3\}$ is not among those selected by μ with positive probability.

Theorem 2. ([7]) *Consider two independent random Cantor sets C_1 and C_2 whose joint survival distributions μ and λ satisfy* (6)*, the joint survival condition.*

(a) *If $\gamma_k > 1$ for all $k \in \mathbb{A}$, then $C_1 - C_2$ contains an interval a.s. on $\{C_1 - C_2 \neq \emptyset\}$.*
(b) *If there exists a $k \in \mathbb{A}$ with $\gamma_k, \gamma_{k+1} < 1$, then $C_1 - C_2$ contains no interval a.s.*

Proof. A proof for joint survival distributions corresponding to M independent Bernoulli variables is given in [7]. An extension of this proof for general survival distributions satisfying the joint survival condition is given in [10]. □

Theorem 2 is in general not conclusive for all pairs $(\mathbf{p}, \mathbf{q})$. Some γ_k might be larger than 1 while others are smaller than 1, though never two consecutively. Another gap, which is not uncommon in branching process theorems, has to do with the boundary case $\gamma_k = 1$. Consider for example the symmetric $M = 3$ case: in this case we always have $\gamma_0 \geq \gamma_1 = \gamma_2$, so it follows that the theorem is conclusive for both $\gamma_1 > 1$ and $\gamma_1 < 1$; it is still not conclusive, however, for the boundary case $\gamma_1 = 1$.

Mora, Simon and Solomyak ([22]) have shown that the condition $\gamma_0\gamma_1 \cdots \gamma_{M-1} > 1$ (cf. the orthogonal projection case), with a small extra requirement, implies that the Lebesgue measure of $C_1 - C_2$ is positive almost surely. With this they can show that there exist random Cantor sets whose difference does not contain an interval, but has positive Lebesgue measure.

The scope of Theorem 2 can be lifted by using the idea of higher-order Cantor sets.

4.3. Higher-order Cantor sets

We already considered order 2 random Cantor sets in Section 2.4. Essentially, the n^{th}-*order random Cantor set* is constructed by 'collapsing' n steps of this construction into one step.

We denote all entities of an n^{th}-order random Cantor set with a superscript ${}^{(n)}$.

Formally, the n^{th}-order random Cantor set ($n \geq 0$) of an M-adic random Cantor set with joint survival distribution μ is an M^n-adic random Cantor set with the joint survival distribution $\mu^{(n)}$ defined below. Note that the alphabet $\mathbb{A}^{(n)}$ is the set $\{0, \dots, M^n - 1\}$. The joint survival distribution $\mu^{(n)} : 2^{\mathbb{A}^{(n)}} \to [0,1]$ is determined uniquely by requiring that

$$X_k^{(n)} \sim \prod_{i=1}^{n} X_{k_1 \dots k_i} = X_{k_1} X_{k_1 k_2} \cdots X_{k_1 \dots k_n},$$

for all $k = k_1 M^{n-1} + \cdots + k_{n-1} M + k_n \in \mathbb{A}^{(n)}$, where the $X_{k_1 \dots k_i}$ are defined in (2), and in particular

$$\Big\{ k_{m+1} \in \mathbb{A}^{(n)} : X^{(n)}_{k_1 \dots k_m k_{m+1}} = 1 \Big\} \sim \mu^{(n)}.$$

for all $k_1 \dots k_m \in \mathcal{T}^{(n)}$. From this definition, it is clear that the higher-order marginal probabilities are given by

$$p_k^{(n)} := \mathbb{P}_{\mu^{(n)}} \left(X_k^{(n)} = 1 \right) = \prod_{i=1}^{n} \mathbb{P}_\mu \left(X_{k_1 \dots k_i} = 1 \right) = \prod_{i=1}^{n} p_{k_i}, \tag{7}$$

for all $k = k_1 M^{n-1} + \cdots + k_n \in \mathbb{A}^{(n)}$, because $X_{k_1}, X_{k_1 k_2}, \dots, X_{k_1 \dots k_n}$ are independent.

The joint survival condition (6) nicely propagates to higher-order Cantor sets. If μ satisfies the joint survival condition, then

$$\mathrm{Supp_m}(\mu^{(n)}) = \Big\{ k = k_1 M^{n-1} + \cdots + k_n \in \mathbb{A}^{(n)} : p_k^{(n)} = \prod_{i=1}^{n} p_{k_i} > 0 \Big\},$$

which implies that $\mu^{(n)}$ satisfies the joint survival condition as well:

$$\mu^{(n)} \left(\mathrm{Supp_m}(\mu^{(n)}) \right) = \left(\mu \left(\mathrm{Supp_m}(\mu) \right) \right)^{1 + a + a^2 + \cdots + a^{n-1}} > 0,$$

where $a := \mathrm{Card}(\mathrm{Supp_m}(\mu))$. This ensures that Theorem 2 can be successfully applied to higher-order Cantor sets $C_1^{(n)}$ and $C_2^{(n)}$ when the joint survival condition holds for μ and λ.

The key observation regarding higher-order Cantor sets is that for all $n \geq 1$

$$C^{(n)} \sim \bigcap_{m=1}^{\infty} C^{n \cdot m} = C, \tag{8}$$

hence statements such as Theorem 2 can be applied to higher-order correlation coefficients $\gamma_k^{(n)}$ in order to get results not only for $C_1^{(n)} - C_2^{(n)}$, but for $C_1 - C_2$ as well. This approach will lead to Theorem 3 in the next section.

4.4. Lower spectral radii

The following definition due to Gurvits ([16]) generalizes the concept of 'spectral radius' to a *set* of matrices.

Let $\|\cdot\|$ be a submultiplicative norm on $\mathbb{R}^{d\times d}$ and $\Sigma \subseteq \mathbb{R}^{d\times d}$ a finite non-empty set of square matrices. The *lower* and *upper spectral radius* of Σ are defined by

$$\underline{\rho}(\Sigma) := \liminf_{n\to\infty} \underline{\rho}_n(\Sigma, \|\cdot\|), \qquad \underline{\rho}_n(\Sigma, \|\cdot\|) := \min_{A_1,\dots,A_n\in\Sigma} \|A_1\cdots A_n\|^{1/n}, \tag{9}$$

$$\bar{\rho}(\Sigma) := \limsup_{n\to\infty} \bar{\rho}_n(\Sigma, \|\cdot\|), \qquad \bar{\rho}_n(\Sigma, \|\cdot\|) := \max_{A_1,\dots,A_n\in\Sigma} \|A_1\cdots A_n\|^{1/n}. \tag{10}$$

Note that for $\Sigma = \{A\}$, with A a square matrix, the lower and upper spectral radius are equal to $\rho(A)$, the spectral radius of A.

For the algebraic difference of two M-adic random Cantor sets, we are interested in the lower spectral radius of a set of matrices

$$\Sigma_{\mathcal{M}} := \{\mathcal{M}(0), \dots, \mathcal{M}(M-1)\}. \tag{11}$$

To define these matrices we first introduce real numbers m_k for $k = -M+1, -M+2, \dots, M-2, M-1$ by

$$m_k = \sum_{i,j\in\mathbb{A}:\, i-j=k} p_i q_j, \tag{12}$$

and put $m_{-M} = m_M = 0$. The number m_k is the expected number of squares in the tilted $C_1^1 \times C_2^1$ whose right half-projects on the interval $\frac{\sqrt{2}}{2} I_k$, see Figure 6 for an illustration.

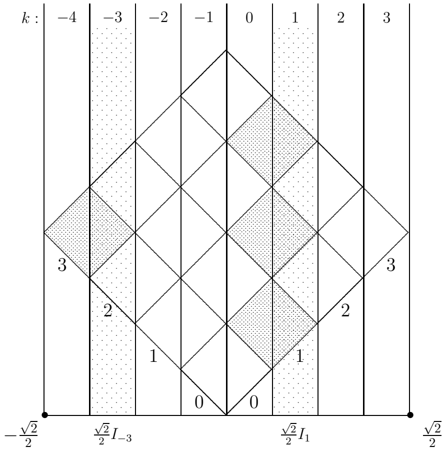

FIGURE 6. When $M = 4$, $m_1 = p_1 q_0 + p_2 q_1 + p_3 q_2$, and $m_{-3} = p_0 q_3$.

Next we define the so-called *growth* matrices $\mathcal{M}\,(k)$ by

$$\mathcal{M}\,(k) = \begin{bmatrix} m_{k+1-M} & m_{k-M} \\ m_{k+1} & m_k \end{bmatrix}, \tag{13}$$

for $k \in \mathbb{A}$. The m_k's are related to the γ_k's by

$$m_k + m_{k-M} = \gamma_k.$$

The following theorem provides a generalization of Theorem 2 for the marginally symmetric case $\mathbf{p} = \mathbf{q}$ (where $m_{-k} = m_k$), using the concept of the lower spectral radius.

Theorem 3. ([10]) *Consider the algebraic difference $C_1 - C_2$ of two M-adic independent random Cantor sets C_1 and C_2 whose joint survival distributions μ and λ satisfy the joint survival condition* (6) *and have equal vectors of marginal probabilities $\mathbf{p} = \mathbf{q}$ such that $m_k > 0$ for all $k \in \mathbb{A}$. Let $\Sigma_{\mathcal{M}}$ be as in* (11)*.*

(a) *If $\underline{\rho}(\Sigma_{\mathcal{M}}) > 1$, then $C_1 - C_2$ contains an interval a.s. on $\{C_1 - C_2 \neq \emptyset\}$.*
(b) *If $\underline{\rho}(\Sigma_{\mathcal{M}}) < 1$, then $C_1 - C_2$ contains no intervals a.s.*

It is not known whether a similar result holds for the marginally asymmetric case $\mathbf{p} \neq \mathbf{q}$. It is clear however, that the joint survival condition is too strong: there are families of μ's who do not satisfy the joint survival condition for which Theorem 3 is true. An improvement of this condition has recently been given in [11].

Note that the interval or not question is entirely determined by the marginal vector $\mathbf{p}$, and that Theorem 3 yields a surface in $\mathbb{R}^M$ that separates the interval case from the no interval case. For $M = 2$ and $M = 3$ this surface has been determined (in respectively [10] and [7]). For all $M \geq 4$ the problem is open. This is connected to the fact that in general it is very hard to determine a lower spectral radius (see [26]).

5. Larsson's random Cantor sets

Let a and b be two positive numbers with

$$a > \frac{1}{4} \quad \text{and} \quad 3a < 1 - 2b. \tag{14}$$

Larsson's random Cantor set is obtained as follows: first remove an interval of length a from the middle of the unit interval, and intervals of length b from both ends. Then put intervals of length a according to a uniform distribution in the remaining two parts $\left[b, \frac{1}{2} - \frac{a}{2}\right]$ and $\left[\frac{1}{2} + \frac{a}{2}, 1 - b\right]$ – note that this is possible by the second part of condition (14). These two randomly chosen intervals of length a are called the level 1 intervals of the random Cantor set $C_{a,b}$. We write $C^1_{a,b}$ for their union. In both of the two level 1 intervals we repeat the same construction (scaled by a) independently of each other and of the previous step. In this way we obtain four disjoint intervals of length a^2. Similarly, the set $C^n_{a,b}$ consists of 2^n

level n intervals of length a^n. Then Larsson's random Cantor set is defined by (see Figure 7 for an illustration)

$$C_{a,b} := \bigcap_{n=1}^{\infty} C_{a,b}^n.$$

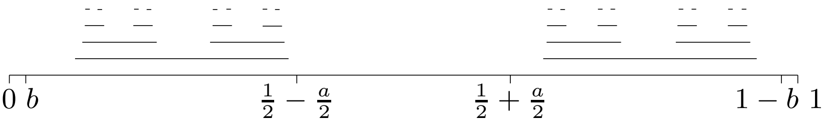

FIGURE 7. The construction of the Cantor set $C_{a,b}$. The figure shows $C_{a,b}^1, \dots, C_{a,b}^4$.

The next theorem was stated by P. Larsson.

Theorem 4. ([19],[8]) *Let a and b satisfy* (14)*, and let C_1, C_2 be independent random Cantor sets having the same distribution as $C_{a,b}$ defined above. Then the algebraic difference $C_1 - C_2$ almost surely contains an interval.*

Note that the first part of condition (14) implies that (surely!)

$$\dim_H(C_{a,b}) = -\frac{\log 2}{\log a} > \frac{1}{2}.$$

Theorem 4 hence implies that the Palis-Takens conjecture holds for the Larsson random Cantor sets.

Larsson's proof is based on a limit theorem for multitype branching processes with continuous state space ([17]), but both in his paper in the Comptes Rendues ([19]), and in his thesis ([20]) he fails to justify the use of this limit theorem. For a correct proof see [8].

6. Final remarks

The random Cantor sets considered here form a very natural class of random fractals. In fact, they have a relation to the most natural random fractal: Brownian motion. Yuval Peres ([24]) has shown that for $d \geq 3$ the path of Brownian motion (restricted to the unit cube) 'sees' sets in the same way as fractal percolation with parameter $p = 2^{2-d}$. This means that there exist positive constants a and b such that for any Borel set B

$$a\mathbb{P}\left(B \cap C \neq \emptyset\right) \leq \mathbb{P}\left(B \cap W((0,\infty]) \neq \emptyset\right) \leq b\mathbb{P}\left(B \cap C \neq \emptyset\right),$$

where C is d-dimensional fractal percolation with parameter 2^{2-d}, and $W((0,\infty])$ is the path of Brownian motion.

The class of M-adic Cantor sets considered here can be considerably extended by letting the survival of a subinterval depend on the presence/absence of the neighbouring intervals, see [9]. Figure 8 illustrates one of the new phenomena that may occur.

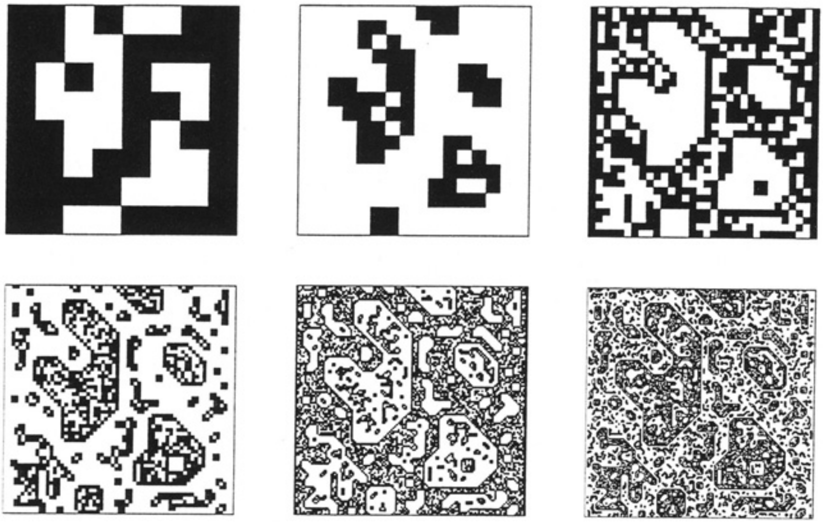

FIGURE 8. Six iterates of an example of fractal percolation with neighbour interaction.

Acknowledgements

I am grateful to Bram Kuijvenhoven for textual contributions to this paper. Also Figures 1 and 5 are due to him. Figure 7 is due to Balázs Székely, Figure 8 to Peter van der Wal. I thank Andreas Wust and Henk Don for carefully reading the paper, and their useful comments.

References

[1] Carlos A. Cabrelli, Kathryn E. Hare, and Ursula M. Molter. Sums of Cantor sets. *Ergodic Theory Dynam. Systems*, 17(6):1299–1313, 1997.

[2] Carlos A. Cabrelli, Kathryn E. Hare, and Ursula M. Molter. Sums of Cantor sets yielding an interval. *J. Aust. Math. Soc.*, 73(3):405–418, 2002.

[3] Lincoln Chayes. Aspects of the fractal percolation process. In *Fractal geometry and stochastics* (*Finsterbergen,* 1994), volume 37 of *Progr. Probab.*, pages 113–143. Birkhäuser, Basel, 1995.

[4] C.G.T. de A. Moreira and J.-C. Yoccoz. Stable intersections of regular Cantor sets with large Hausdorff dimensions. *Ann. of Math.* (2), 154:45–96, 2001.

[5] F.M. Dekking and G.R. Grimmett. Superbranching processes and projections of random Cantor sets. *Probab. Theory Related Fields*, 78(3):335–355, 1988.

[6] F.M. Dekking and R.W.J. Meester. On the structure of Mandelbrot's percolation process and other random Cantor sets. *J. Statist. Phys.*, 58(5-6):1109–1126, 1990.

[7] F.M. Dekking and K. Simon. On the size of the algebraic difference of two random Cantor sets. *Random Structures Algorithms*, 32(2):205–222, 2008.

[8] F.M. Dekking, K. Simon, and B. Szekely. The algebraic difference of two random Cantor sets: the Larsson family. `http://arxiv.org/abs/0901.3304`, 2009.

[9] F.M. Dekking and P. v.d. Wal. Fractal percolation and branching cellular automata. *Probab. Theory Related Fields*, 120(2):277–308, 2001.

[10] F. Michel Dekking and Bram Kuijvenhoven. Differences of random Cantor sets and lower spectral radii. `http://arxiv.org/abs/0811.0525`, 2008.

[11] Henk Don. *The Four gap theorem and Differences of random Cantor sets*. Delft University of Technology, Delft, January 2009. MSc Thesis, adviser Michel Dekking.

[12] Kemal Ilgar Eroğlu. On the arithmetic sums of Cantor sets. *Nonlinearity*, 20(5):1145–1161, 2007.

[13] K.J. Falconer. Random fractals. *Math. Proc. Cambridge Philos. Soc.*, 100(3):559–582, 1986.

[14] K.J. Falconer and G.R. Grimmett. On the geometry of random Cantor sets and fractal percolation. *J. Theoret. Probab.*, 5(3):465–485, 1992.

[15] K.J. Falconer and G.R. Grimmett. Correction: "On the geometry of random Cantor sets and fractal percolation" [J. Theoret. Probab. **5** (1992), no. 3, 465–485. *J. Theoret. Probab.*, 7(1):209–210, 1994.

[16] Leonid Gurvits. Stability of discrete linear inclusion. *Linear Algebra Appl.*, 231:47–85, 1995.

[17] Theodore E. Harris. *The theory of branching processes*. Dover Phoenix Editions. Springer, Berlin, 1963. Corrected reprint of the 1963 original [Springer, Berlin; MR0163361 (29 #664)].

[18] J.-P. Kahane and J. Peyrière. Sur certaines martingales de Benoit Mandelbrot. *Advances in Math.*, 22(2):131–145, 1976.

[19] Per Larsson. L'ensemble différence de deux ensembles de Cantor aléatoires. *C.R. Acad. Sci. Paris Sér. I Math.*, 310(10):735–738, 1990.

[20] Per Larsson. *The difference set of two Cantor sets*, volume 11 of *U.U.D.M. Reports*. Uppsala University, Thunsbergsvägen 3, 752 38 Uppsala, Sweden, May 1991. Thesis, adviser Lennart Carleson.

[21] Benoit B. Mandelbrot. *The fractal geometry of nature*. W.H. Freeman and Co., San Francisco, Calif., 1982. Schriftenreihe für den Referenten. [Series for the Referee].

[22] Peter Mora, Karoly Simon, and Boris Solomyak. The Lebesgue measure of the algebraic difference of two random Cantor sets. *Preprint*, 2008.

[23] Jacob Palis and Floris Takens. *Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations*, volume 35 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1993. Fractal dimensions and infinitely many attractors.

[24] Yuval Peres. Intersection-equivalence of Brownian paths and certain branching processes. *Comm. Math. Phys.*, 177(2):417–434, 1996.

[25] Yuval Peres and Pablo Shmerkin. Resonance between Cantor sets. *Ergodic Theory and Dynamical Systems*, 29:201–221, 2009.

[26] John N. Tsitsiklis and Vincent D. Blondel. The Lyapunov exponent and joint spectral radius of pairs of matrices are hard – when not impossible – to compute and to approximate. *Math. Control Signals Systems*, 10(1):31–40, 1997.

Michel Dekking
Delft Institute of Applied Mathematics
Technical University of Delft
The Netherlands
e-mail: `F.M.Dekking@tudelft.nl`

Appendix

Fractal Geometry and Stochastics I

Birkhäuser, Progress in Probab. 37, 1995

CONTENTS

Fractal Geometry and Stochastics II

Birkhäuser, Progress in Probab. 46, 2000

CONTENTS

Fractal Geometry and Stochastics III

Birkhäuser, Progress in Probab. 57, 2004

CONTENTS

1. Fractal Sets and Measures

2. Fractals and Dynamical Systems

3. Stochastic Processes and Random Fractals

4. Fractal Analysis in Euclidean Space

5. Harmonic Analysis on Fractals